On the use of GSV-SVM for Speaker Diarization and Tracking

Viet Bac Le, Claude Barras, Marc Ferràs

Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

On the use of GSV-SVM for Speaker Diarization and Tracking

Viet Bac Le[1], *Claude Barras*[1,2] *and Marc Ferràs*[1,*]

[1]LIMSI-CNRS, BP 133, 91403, Orsay, France
[2]Université Paris-Sud, F-91405, Orsay, France
{levb,barras}@limsi.fr, ferras@furui.cs.titech.ac.jp

Abstract

In this paper, we present the use of Gaussian Supervectors with Support Vector Machines classifiers (GSV-SVM) in an acoustic speaker diarization and a speaker tracking system, compared with a standard Gaussian Mixture Model system based on adapted Universal Background Models (GMM-UBM). GSV-SVM systems (which share the adaptation step with the GMM-UBM systems) are observed to have comparable performances: for acoustic speaker diarization, the GMM-UBM system outperforms the GSV-SVM system on ESTER2 data but the latter system works better in the speaker tracking system. In particular, the linear combination of two systems at the score level outperforms each individual system.

1. Introduction

With the introduction of Support Vector Machines (SVM) into the speaker recognition community in recent years, much research has focused on finding novel speaker-relevant and robust features. SVMs are capable of working on very high dimensional, even under-sampled, data without losing their generalization ability. Gaussian Supervectors (GSV-SVM) [1] is one of such approaches, which successfully combines, via MAP adaptation, both GMM and SVM modeling together in a simple and easy-to-develop framework. In a different direction, Maximum-Likelihood Linear Regression (MLLR) and Constrained MLLR (CMLLR) have been recently used for feature extraction in SVM-based speaker recognition [2, 3]. GSV-SVM and (C)MLLR-SVM techniques become more and more present in state-of-the-art text-independent speaker recognition systems.

In the context of the Quaero program [1], we aim at improving the state-of-the-art in automatic audio-visual document structuring and indexing. We currently work on speaker diarization and speaker tracking for a various types of audio data: broadcast news, broadcast conversational speech, web contents and talk-shows. Since the use of a GMM-based speaker identification stage in a multi-stage speaker diarization system has been demonstrated to perform much better than the initial single-stage BIC clustering system [4], we are testing the integration of SVM-based speaker identification techniques in this system. Similarly, these SVM techniques are being investigated in the speaker verification step of a speaker tracking system.

The remainder of this paper is organized as follows: Section 2 describes the experimental data used in both speaker diarization and tracking. The use of the GSV-SVM technique and the combination of the GSV-SVM and GMM-UBM for diarization and tracking are shown in section 3 and 4, respectively. Section 5 concludes the work and gives some future work.

This work has been partially financed by OSEO under the Quaero program

*Marc Ferràs is now with the Furui Laboratory - Tokyo Institute of Technology (http://www.furui.cs.titech.ac.jp)

[1]http://www.quaero.org

2. Data

The radio broadcast data recorded and annotated for the Speaker Diarization and Speaker Tracking tasks of the French ESTER-2 Evaluation Campaign [5] are used in these experiments.

For speaker tracking, the training data contains 111 radio shows recorded in 1999-2003 from France Inter (30 shows - 26 hours), RFI (68 shows - 69 hours) and African n^o1 (13 shows - 10 hours) radio stations for a total of about 105 hours. A list of 115 target speakers (83 male and 32 female) with gender and radio source was provided. For each target speaker from this list, all target segments were extracted from the ESTER2 training data and at least 1 minute of speech per speaker is assured.

The development data (ESTER2-Dev) contains 20 radio shows recorded in July, 2007 from France Inter (5 shows - 2 hours), RFI (2 shows - 40 min.), Africa n^o1 (9 shows - 2 hours 20 min.) and TVME (2 shows - 1 hours) radio stations for a total of about 6 hours. The evaluation data (ESTER2-Eval) contain 26 radio shows recorded in Jan-Feb 2008 from RFI (7 shows - 1 hours 10 min.), France Inter (6 shows - 3 hours 25 mins), TVME (4 shows - 1 hour 10 min.) and Africa n^o1 (9 shows - 1 hour 30 min.) for a total of about 7 hours. Some statistics on the number of speakers per show, speaking time per segment and per speaker are shown in Table 1.

The impostor speakers set for both SVM-based speaker diarization and SVM-based speaker tracking are selected from the ESTER-1 data [6] and then excluded from the list of target speakers. A total of 385 impostors (263 males and 122 females) are classified by gender and channel condition: 72 male-telephones (male speakers spoken in the telephone channel), 191 male-studios, 18 female-telephones and 104 female-studios.

3. Speaker Recognition Techniques for Acoustic Speaker Diarization

The acoustic speaker diarization system is based on the 2005 LIMSI multi-stage speaker diarization system which was developed for NIST RT-04F evaluation on English broadcast news data and presented in 2005 for ESTER-1 evaluation campaign on French radio broadcast news [4].

In general, this system performs the following steps:

- Feature extraction of PLP-like Mel frequency cepstral coefficients,
- Speech Activity Detection with Viterbi decoding using Gaussian Mixture Models of speech and non speech,

Table 1: Some statistics (min, max, mean and standard deviation) on the number of speakers per show, speaking time per segment and per speaker (in second)

	ESTER2-Dev		ESTER2-Eval	
	min - max	μ / σ	min - max	μ / σ
# Speaker	9 - 25	16 / 4.4	5 - 26	11.5 / 5.9
Spk len (s)	0.4 - 630	65 / 96	0.5 - 903	80 / 128
Seg len (s)	0.3 - 193	17 / 23	0.2 - 145	14 / 20.6

- Segmentation into acoustically homogeneous small segments using a Gaussian divergence measure,
- First agglomerative BIC speaker clustering stage with single Gaussian models associated to the *BIC* criterion,
- Second SID speaker clustering stage relying on Cross Log-likelihood Ratio (*CLR*) between more complex speaker models MAP-adapted from a Universal Background Model (*UBM*).

We note that in the SID clustering stage of the system, standard speaker recognition techniques based on GMM-UBM or GSV-SVM models are used. In the following, we present the use of speaker recognition techniques involved in our speaker diarization system.

3.1. GMM-UBM system

Acoustic features are extracted from the speech signal on the 0-8kHz bandwidth for studio speech segments and 0-3.8kHz for telephone speech segments every 10ms using a 30ms window. The feature vector consists of 15 PLP-like cepstrum coefficients computed on a Mel frequency scale plus 15 delta coefficients and delta energy, for a total of 31 features. Feature warping normalization [7], which reshapes the short-term histogram of the cepstral coefficients into a Gaussian distribution is performed using a sliding window of 3 seconds in order to reduce the effect of the acoustic environment.

For each gender and channel condition (studio, telephone) combination, a Multilingual Universal Background Model (*UBM*) [8] with 128 diagonal Gaussians was trained on a Multilingual Broadcast Corpus which contains broadcast data in Arabic, Chinese, English, French, Italian, Russian and Spanish. Then, for each speaker cluster c_i, speaker model λ_i is derived by MAP adapting the channel and gender matched *UBM*'s parameters using the acoustic frames X_i belong to the cluster c_i.

An agglomerative clustering is performed separately for each gender and bandwidth condition using the Cross Log-likelihood Ratio (*CLR*) which was defined as follow:

$$CLR(\lambda_i, \lambda_j) = \frac{1}{L_i} log \frac{f(X_i|\lambda_j)}{f(X_i|\lambda_{UBM})} + \frac{1}{L_j} log \frac{f(X_j|\lambda_i)}{f(X_j|\lambda_{UBM})}$$

where $f(X_i|\lambda)$ is the likelihood of the acoustic frames X_i given the model λ and L_i is the number of frame of X_i. *CLR* is thus a symmetric measure.

After each clustering iteration, the two nearest clusters c_i and c_j (having the highest *CLR* score) are merged and a new model is estimated for cluster c_{ij}. The clustering stops when the highest *CLR* score between all clusters is below a given threshold (estimated on a development dataset).

3.2. GSV-SVM system

By integrating the GSV-SVM technique into the SID clustering stage of the diarization system, we replace the computation of the *CLR* score (given above) by the score calculated from the GSV-SVM. The GSV-SVM system uses Gaussian mean supervectors of a GMM as features. GMMs are adapted using MAP adaptation from gender and channel dependent UBMs (same UBMs with the GMM-UBM system). We use 128 Gaussians and variance-normalization. Then, speaker-related features are obtained by stacking and taking the difference of mean supervectors of the adapted GMM model with the UBM model. SVM training is performed with a linear kernel using SVMTorch [9] from IDIAP.

3.3. Score-Level System Combination

Combining two or several diarization systems may improve the performance over the best individual system. For instance, several combining strategies were presented in [10] using a "piped" system (use the output of the first system as input of the second system) or by merging the proposed segmentations outputted by different systems. A 'cluster voting' technique was also described in [11] which maintains areas of agreement and voting using confidences or an external judging scheme in areas of conflict.

In our experiment, the combination of the GSV-SVM and GMM-UBM systems is performed at score-level during clustering rather than on the output of the individual systems. During the SID clustering process, a weighted average score is computed from the GSV-SVM and GMM-UBM scores. The combination weight is optimized on the development data.

3.4. Experimental results

The speaker diarization task performance is measured via an optimum one-to-one mapping between the reference speakers and the hypothesis speakers. The primary performance measure for speaker diarization task, referred to as the speaker match error (or speaker error), is the fraction of speaker time that is not attributed to the correct speaker, given the optimum speaker mapping. Another measure is the overall speaker diarization error rate (DER) which includes the missed and false alarm speaker times, thus taking speech/non-speech detection errors into account [12]. For additional analysis, cluster purity and coverage errors are also provided [4].

For each diarization system, speaker match error is optimized on ESTER2-Dev data and a clustering threshold is chosen and applied to the ESTER2-Eval data.

Figure 1 shows some results of the GMM-UBM, GSV-SVM systems and the linear combination of two systems as a function of the combination weight. We observe that when the combination weight is varied from 0.1 to 0.9, the DERs of the combined systems are correlatively changed on the development and the evaluation data. The combination weight couple of (0.8, 0.2) seems to be the optimal values for ESTER-2 data.

More performance measures are detailed in Table 2. Although the GSV-SVM system never outperforms the GMM-UBM system on both ESTER2-Dev and ESTER2-Eval data, the linear combination of the two systems overcomes the best individual system (GMM-UBM) by 8.8% and 13.0% relative on the development and evaluation data, respectively.

Table 2: Results of speaker diarization for GMM-UBM, GSV-SVM and their system combination on ESTER2 data

System	clustering threshold	ESTER2-Dev				ESTER2-Eval			
		%purity error	%coverage error	%spk error	%DER	%purity error	%coverage error	%spk error	%DER
GMM-UBM	1.65	5.4	5.8	9.4	11.07	2.9	6.1	7.8	9.57
GSV-SVM	1.62	5.0	7.2	11.0	12.67	4.4	7.7	10.7	12.52
GMM(0.8)+SVM(0.2)	1.35	5.8	4.6	**8.5**	**10.10**	3.1	4.8	**6.5**	**8.33**

Figure 1: Score-level GMM-UBM and GSV-SVM diarization system combination with different combination weights

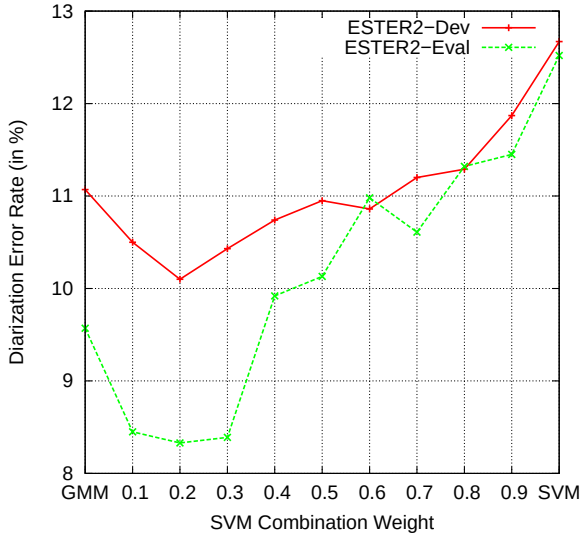
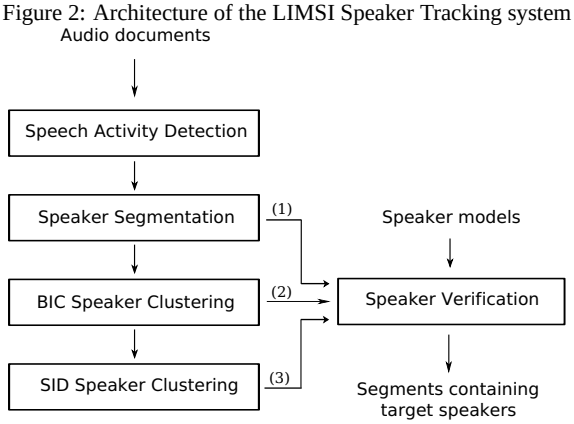


Figure 2: Architecture of the LIMSI Speaker Tracking system



4. Combining GMM-UBM and GSV-SVM for Speaker Tracking

4.1. Task Definition

Speaker tracking aims at detecting regions of a spoken document uttered by a given speaker. A list of target speakers is pre-defined. In the reported experiments on ESTER2 evaluation data, target speakers are journalists or politicians, for example.

4.2. System Architecture

Figure 2 presents the architecture of the LIMSI Speaker Tracking system. It is structured in two principal parts: speaker segmentation (or acoustic speaker diarization) and speaker verification.

For an input audio document, the LIMSI GMM-based multi-stage acoustic speaker diarization system is firstly used for segmenting speech data into homogeneous segments and clustering these segments by speaker.

Then for speaker verification, each speech segment (branch (1) in figure 2) or cluster of speech segments uttered by the same speaker (branch (2) and (3) in figure 2) are scored against all targets in parallel and the target model with the highest score is taken. If this score is bigger than a pre-defined decision threshold, the test segment (or speaker cluster) is labeled with the name of the matched target speaker. For system implementation, the LIMSI speaker identification system as proposed to CLEAR'07 evaluation [13] was initially used. Both GMM-UBM and GSV-SVM are actually investigated for speaker verification. We note that no channel compensation (Factor Analysis or Nuisance Attribute Projection, ...) has been performed in our experiments, since this was not felt critical on the ESTER2 data where most target speakers (journalists and politicians) was recorded in the similar conditions (studio or telephone).

4.3. GMM-UBM system

Acoustic features and the UBMs used for speaker verification are the same as for speaker diarization.

For each target speaker, a speaker-specific GMM is trained by Maximum A Posteriori (MAP) adaptation [14] of the means of the matching UBM. We note that, for a target speaker, we may have two different GMM models trained in different channel condition (studio, telephone). Target models are MAP adapted using 3 iteration of the EM algorithm and a prior factor $\tau = 10$. The GMM-UBM approach has proved to be very successful for text-independent speaker recognition, since it allows for robust estimation of the target models even with a limited amount of enrollment data [8].

During the verification phase, each test segment or speaker cluster X is scored against both the target model λ_k and the UBM model in the same gender and channel condition. For a given test segment X and a target model λ_k, the decision score $S(X|\lambda_k)$ is a log-likelihood ratio:

$$S(X|\lambda_k) = \frac{1}{L_X}\left[\log f(X|\lambda_k) - \log f(X|\lambda_{UBM})\right]$$

where $f(X|\lambda_k)$ is the likelihood of the speech segment X of L_k frames for a given model λ_k and $f(X|\lambda_{UBM})$ is the likelihood of X for a gender- and channel-matching UBM. The target model with the highest log-likelihood ratio is chosen:

$k^* = \text{argmax}_k S(X|\lambda_k)$. A pre-defined decision threshold is applied on the selected target model. The decision threshold is the same for all the target speaker models for both telephone and studio condition. In order to accelerate the GMMs computation, top-Gausian scoring was used, restricting the log-likelihood estimation to the 10 top scoring out of 128 components of the UBM for each frame [8].

4.4. GSV-SVM system

For GSV-SVM based speaker verification, the same system prototype as the SID clustering stage of the acoustic speaker diarization system have been used. For each target speaker, a speaker-specific GMM is created using MAP adaptation from gender and channel dependent 128-Gaussians UBMs. A speaker-related feature vector (GSV) is obtained by taking the difference of mean supervectors of the adapted GMM model with the UBM model. SVM training is also performed with a linear kernel using SVMTorch.

During the verification phase, each test segment or speaker cluster is scored against consecutively impostors and against the true speaker matching the gender and channel condition.

For system combination, the GMM-UBM and GSV-SVM systems are also combined at the score level: a weighted average score is computed from the GSV-SVM and GMM-UBM scores. This system combination method is similar with the score-level fusion technique used in speaker recognition [3].

4.5. Performance measures

The speaker tracking performance is measured by recall (RCL), precision (PRC), F-measure (F) and mean F-measure (mean-F) which are frequently used for evaluating the performance of information retrieval systems. They are defined as follow:

$$RCL = \frac{\text{Correctly detected target speaker time}}{\text{Reference target speaker time}}$$

$$PRC = \frac{\text{Correctly detected target speaker time}}{\text{Hypothesized target speaker time}}$$

$$F = \frac{2 \times RCL \times PRC}{RCL + PRC}$$

$$\text{mean-}F = \frac{1}{N_{spk}} \cdot \sum_{i=1}^{N_{spk}} F_i$$

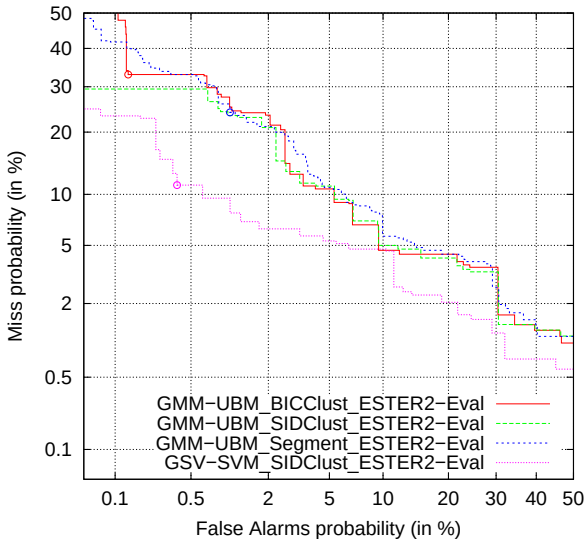
where N_{spk} is number of target speakers.

To evaluate these measure, we use the evaluation tool proposed for the tracking task of the ESTER-2 Evaluation Campaign [5] Thus, for each system, the decision threshold have been selected by maximizing the global F-measure on ESTER2-Dev data.

4.6. Experimental Results

Figure 3 shows the DET curves evaluated on ESTER2-Eval data using different type of test segments outputted from the diarization system (see figure 2): (1) a speech segment, (2) a speaker cluster (i.e. all speech segments uttered by same clustered speaker are tested together) resulting from the BIC clustering and (3) a speaker cluster resulting from the SID clustering.

Figure 3: DET curve on ESTER2-Eval data for GMM-UBM and GSV-SVM systems (with different segmentation methods for GMM-UBM)



Only the latter one is used for GSV-SVM system. Circles are drawn at minimal DCF operating points (we used the conventional cost function for speaker identification). To draw a DET curve, the duration of the test segment is taken into account (target and non-target scores are extracted every 3 seconds). It is interesting to observe that the GSV-SVM system (when using SID speaker clustering) performs significantly better than GMM-UBM systems.

Table 3 shows some results of recall rate, precision rate, F-measure and mean-F evaluated on ESTER2-Dev and ESTER2-Eval data. The first three lines compare the GMM-UBM systems using different type of test segments. We observed that speaker tracking performs better when speaker clustering (BIC or SID) is performed.

Moreover, the GSV-SVM system works significantly better than the GMM-UBM system on ESTER2-Eval data although that the performance of two systems are similar on the development data. Anyway, the linear combination of two systems outperforms the best individual system. The best combination weight (optimized on ESTER2-Dev data) is 0.6 for GMM-UBM system and 0.4 for GSV-SVM system.

5. Conclusions and future work

The use of GSV-SVM technique in the speaker diarization and tracking systems was proposed in this paper. GSV-SVM systems (which share the GMM adaptation step with the GMM-UBM systems) are observed to have performance comparable to the GMM-UBM ones. Especially, the linear combination of two types of systems works significantly better than the individual systems.

We believe the number of impostor speakers used in these experiments to be too small (especially in the telephone channel) to build a good SVM classifier. A bigger (monolingual or multilingual) impostor dataset needs to be investigated in the future. Other SVM features (like CMLLR, MLLR or Lattice MLLR [15]) could also be investigated.

Table 3: Recall, Precision, F-measure and mean F-measure of speaker tracking for GMM-UBM, GSV-SVM and their system combination on ESTER2-Dev and ESTER2-Eval data

System	Segment or cluster?	ESTER2-Dev				ESTER2-Eval			
		RCL	PRC	F	mean-F	RCL	PRC	F	mean-F
GMM-UBM	Segment	0.387	0.689	0.496	0.727	0.597	0.739	0.660	0.731
GMM-UBM	BIC clust	0.393	0.859	0.540	0.787	0.604	0.748	0.668	0.805
GMM-UBM	SID clust	0.415	0.824	0.552	0.796	0.632	0.755	0.688	0.806
GSV-SVM	SID clust	0.467	0.677	0.553	0.794	0.690	0.785	0.734	0.852
GMM(0.6)+SVM(0.4)	SID clust	0.429	0.849	**0.570**	**0.805**	0.632	0.887	**0.738**	**0.867**

6. References

[1] W. M. Campbell, D.E. Sturim, D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Veriffication", *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[2] A. Stolcke, L. Ferrer, S. Kajarekar, "Improvements in MLLR-Transform-Based Speaker Recognition", *Odyssey'06*, June 2006.

[3] M. Ferràs, C-C. Leung, C. Barras, J-L. Gauvain, "Constrained MLLR for Speaker Recognition", *ICASSP'07*, pages 53-56, Honolulu, Hawaii, April 2007.

[4] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage Speaker Diarization of Broadcast News", *IEEE TASLP*, Vol. 14, No. 5, pp. 1505–1512, September 2006.

[5] S. Galliano, G. Gravier, L. Chaubard, "The ESTER 2 evaluation campaign for the Rich Transcription of French Radio Broadcasts", *Interspeech'09*, Brighton, U.K., 2009.

[6] G. Gravier et al., "The ESTER evaluation campaign of Rich Transcription of French Broadcast News ", *LREC'04*, 2004.

[7] J. Pelecanos and S. Sridharan, "Feature Warping for Speaker Verification," *Proceedings of IEEE Speaker Odyssey*, 2001.

[8] D. Reynolds, T. Quatieri and R. Dunn, "Speaker verification using adapted gaussian mixture models", *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[9] R. Collobert, S. Bengio, "SVMTorch: a Support Vector Machine for Large-Scale Regression and Classification Problems", *Journal of Machine Learning Research*, Vol. 1, pp. 143–160, September 2001.

[10] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre, "The ELISA Consortium Approaches in Broadcast News Speaker Segmentation during the NIST 2003 Rich Transcription Evaluation," *ICASSP'04*, Vol. 1, pp. 37–376, Montreal, May 2004.

[11] S.E. Tranter, "Two-way Cluster Voting to Improve Speaker Diarisation Performance," *ICASSP'05*, Vol. 1, pp. 753–756, Philadelphia, PA, March 2005.

[12] NIST, "Fall 2004 Rich Transcription (RT-04F) Evaluation Plan", August 2004.

[13] C. Barras et al., "The CLEAR'07 LIMSI System for Acoustic Speaker Identification in Seminars", *CLEAR'07*, pp. 233-239, Baltimore, MD, USA, May, 2007.

[14] J.-L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE TSAP*, vol. 2, no. 2, pp. 291-298, April 1994.

[15] M. Ferràs, C. Barras, J-L. Gauvain, "Lattice-based MLLR for speaker recognition", *ICASSP'09*, Taipei, Taiwan, April 2009.