COMPENSATION FOR INTER-FRAME CORRELATIONS IN SPEAKER DIARIZATION AND RECOGNITION

Themos Stafylakis[1,2]*, Patrick Kenny*[1]*, Vishwa Gupta*[1] *and Pierre Dumouchel*[1,2]

[1]Centre de Recherche Informatique de Montreal (CRIM), Canada,
[2]Ecole de Technologie Superieure (ETS), Canada

ABSTRACT

In this paper, we introduce the concept of the effective sample size to speaker diarization and recognition. We show why the use of the nominal sample size is inadequate to feature streams that exhibit inter-frame correlations and how it adversely affects inference. We then discuss the effective sample size, that is the sample size of a set of independent observations that carry the equivalent amount of statistical information about the model parameters and how the scaling factor can be estimated. Our experiments on speaker diarization show that once the effective sample size is adopted, state-of-the-art results can be attained even with single Gaussians and Hierarchical Clustering, and even when the scaling factor is set to be common for all utterances. On speaker recognition, encouraging results are reported on NIST-2010 using iVectors and PLDA.

Index Terms— Speaker recognition, Clustering methods, Bayesian methods

1. INTRODUCTION

Contemporary research in speaker recognition and diarization assumes a Gaussian Mixture Model (GMM) as the generative model of an utterance and focuses on methods to increase the robustness of the estimate of the components' mean values. The iVector representation has been proven to be a very effective way to towards robust estimation, [1]. Having fixed a powerful representation for utterances and speakers, much effort is currently given to quantify the uncertainly in the point-estimates. In [2], a fully Bayesian approach is considered, where the model parameters of PLDA are treated as random variables and therefore integrated out. In [3], the uncertainty in the iVector point-estimate is not discarded, but propagated to the PLDA model. In speaker diarization, where speech segments can be very sort and clusters are of variable duration, several approaches to quantify the uncertainty and use it in order to make a fully or semi-Bayesian treatment of model's order selection possible have been attempted, [4], [5].

When modeling the uncertainty, though, a key-issue that needs to be reconsidered is the independent assumption regarding the front-end features (e.g. MFCC). It is well known and fairly easy to verify that this assumption is false. However, the results of treating them as if they were indeed independent have been underestimated by the speaker recognition and diarization community. Literature in other fields shows that if the nominal sample size is used, the uncertainty of the random variables would be smaller than its actual, [6], [7]. This not only leads to a poor estimate of the posterior expectation of the variable in question (i.e. the point-estimates). More importantly, it causes severe degradation in several model's order selection tasks, including the estimation of number of speakers in diarization, bad calibration of log-likelihood ratios (LLRs) in speaker recognition, and failure of BIC-like criteria in diarization and segmentation. And as we show in the speaker diarization section, there are cases where quantifying the uncertainty properly is at least equally crucial to obtaining complex point-estimates about the true probability distribution of a speaker.

The rest of the paper is organized as follows. In Sect. 2, the concept of the effective sample size (ESS) is introduced along with the issues arising from a naive use of the nominal one. In Sect. 3, a Bayesian approach to speaker diarization is described and experiments are given based on the ETAPE contest. In Sect. 4, the way to incorporate the ESS into the iVector extractor is presented and results are given on NIST2010 data. Finally, conclusions and future work are discussed in Sect. 5, including a formal method to estimate the ESS.

2. INTER-FRAME CORRELATIONS AND EFFECTIVE SAMPLE SIZE

2.1. How inter-frame correlations adversely affect estimators

To gain some intuition about the need of the notion of the ESS let us start by giving an elementary example. Assume the task of estimating the mean μ of a speech segment, in some feature space. A typical feature extractor for speaker and speech recognition (e.g. MFCC) extracts one feature vector every 10ms. From statistical theory, it is well known that the variance $V(\hat{\mu})$ of an estimate $\hat{\mu}$ of μ is reciprocal to the number of observation vectors. Hence, would it be unreasonable to ask why not increasing the frame rate to one feature vector every 1ms and reduce the variance of the estimator by a factor of 10?

The reason why this increase would not be effective is that by doing so, the inter-frame correlation will increase as well. Once a certain frame rate is exceeded, no further decrease in $V(\hat{\mu})$ can be attained, [6]. In order to obtain a fixed frame rate, the speech and speaker communities concluded that the 10ms frame rate is sufficient on average. However, even with the standard frame rate of 10ms, a high degree of inter-frame correlation appears between observation vectors. And unless this correlation is taken into account, several artifacts in estimation and inference would appear.

2.2. The problems with overconfident estimators

2.2.1. The use of BIC in speaker diarization

A domain where a reliable measure the confidence of estimators plays a crucial role is model order selection criteria. Although closed-form expressions will be used in the experiments on diarization, the Bayesian Information Criterion (BIC) serves as a proper introductory example to present the main idea, due to its simplicity, [8], [9]. Consider an audio file of n frames, assume an initial

segmentation into N segments and let two segments be $\mathbf{y}_a$ and $\mathbf{y}_b$, of n_a and n_b sample sizes, respectively. Assume we model each speaker by a single Gaussian, denoted by $\theta = (\mu, \Sigma)$ and having $P = d + d(d+1)/2$ free parameters. The familiar ΔBIC approximation to the log-likelihood ratio (LLR) between the two hypotheses (namely $\mathcal{H}_1$: two different speakers and $\mathcal{H}_0$: a single speaker) is as follows

$$\Delta\text{BIC}_{\mathbf{y}_a,\mathbf{y}_b} = l(\tilde{\theta}_a|\mathbf{y}_a) + l(\tilde{\theta}_b|\mathbf{y}_b) - l(\tilde{\theta}_{a\cup b}|\mathbf{y}_{a\cup b}) - \frac{\lambda}{2}PT(\{n_t\}) \tag{1}$$

In the above expression, $l(\tilde{\theta}_a|\mathbf{y}_a) = n_a\bar{l}(\tilde{\theta}_a|\mathbf{y}_a)$, $\bar{l}(\tilde{\theta}_a|\mathbf{y}_a) = \frac{1}{n_a}\sum_{i=1}^{n_a}\log p(\mathbf{y}_a^{(i)}|\tilde{\theta}_a)$ and $\tilde{\theta}_a$ the MAP estimate of θ_a, although maximum likelihood (ML) estimates $\hat{\theta}_a$ are used more often.
The most common setting is the so-called local-BIC (see [10]), where $T(\{n_t\}) = \log(n_a + n_b)$, although other settings have been proposed, namely the global-BIC, where $T(\{n_t\}) = \log n$ and the segmental-BIC, where $T(\{n_t\}) = \log n_a + \log n_b - \log(n_a + n_b)$, [11].

2.2.2. A consequence of using the nominal sample sizes

It is well known amongst diarization community that ΔBIC with $\lambda = 1$ will hardly merge any segments. Hence, the fudge factor $\lambda > 1$ is placed in order to boost the penalty term over the likelihood ones. However, such a fudge factor lacks of any coherent bayesian interpretation. Using the Laplace approximation to the marginal likelihood, it can be proven that such a fudge factor implies a prior on θ that is sample size dependent, [9]. However, such priors are rejected by the bayesian community.
Let us show why this failure is a results of the use of the nominal sample sizes. Recall that in BIC with ML, the uncertainly in the estimates appears only via (n_a, n_b). For given n_a and n_b, and considering the high inter-frame correlation in speech signals, any statistical test (including ΔBIC) would conclude that the discrepancy between $\hat{\theta}_a$ and $\hat{\theta}_b$ (e.g. Kullback-Leibler divergence, $D_{KL}(\hat{\theta}_a : \hat{\theta}_b)$) is too large for $\mathbf{y}_a$ and $\mathbf{y}_b$ to be regarded as outcomes of a single distribution, $\theta_{a\cup b}$, even in cases where $\mathcal{H}_0$ holds. This is due to the fact that the estimator becomes too confident about $\hat{\theta}_a$ and $\hat{\theta}_b$, due of the use of the nominal sample sizes. In other words, the artificial boost of the penalty term by $\lambda > 1$ aims to compensate for the artificial boost in the precision of the estimator, caused by the use of the nominal sample size.
Therefore, the natural way to tackle this problem is not by boosting the penalty term, but by using the effective sample sizes rn_a and rn_b in the likelihood and penalty terms, where $0 < r \leq 1$ denotes the scaling factor. By doing so, despite the fact that the point-estimates of $\hat{\theta}_a$ and $\hat{\theta}_b$ would remain unchanged (and therefore their discrepancy $D_{KL}(\hat{\theta}_a : \hat{\theta}_b)$), the uncertainty in those estimates would be r^{-1} times higher, making the $\mathcal{H}_0$ hypothesis more plausible.

2.2.3. MAP estimates and posterior expectations

In the case of MAP estimates and/or posterior expectations, the use of n_a instead of rn_a has an additional undesired property. As n_a grows, the observations $\mathbf{y}_a$ overwrite the information carried in the prior of θ_a in a faster rate than the optimal. Hence, in order for the prior to be effective on the posterior, one should use priors that are more sharp around θ_0. However, a prior cannot be arbitrarily sharp (i.e. informative), but it has to reflect our prior beliefs that are ideally based on training data. Therefore, the use of the nominal sample size in correlated data leads to point-estimates on which the actual data overwrites the prior faster than it should. The following experiments with iVectors in speaker recognition and with the use of the prior on the partition space in diarization verify this claim.

3. INCORPORATING THE EFFECTIVE SAMPLE SIZE IN SPEAKER DIARIZATION

In this chapter we show how to take into account the effective sample size into modeling. We will demonstrate how a baseline algorithm and crude modeling of the speaker pdf can become highly competitive, only by using the ESS scaling.

3.1. The hierarchical clustering algorithm for Speaker Diarization

In this section we demonstrate the algorithm that we submitted to ETAPE diarization contest, [13]. This approach, despite the use of a rather crude distribution (a single Gaussian with full-covariance matrix on a 19-dim static-only MFCC space) to model each speaker and a baseline algorithm (Agglomerative Hierarchical Clustering, AHC) was ranked amongst the top of the contest.

3.1.1. Notation and modeling assumptions

The goal is to maximize the posterior of $(\mathbf{s}, K)$ given the data $\mathbf{y} = \{\mathbf{y}^{(i)}\}_{i=1}^n$, where $\mathbf{s} = \{s^{(i)}\}_{i=1}^n, s^{(i)} = 1, 2, \ldots$ is the assignment of frames to speakers and K the number of speakers. The posterior is decomposed as follows

$$P(\mathbf{s}, K|\mathbf{y}, \mathcal{M}) \propto p(\mathbf{y}|\mathbf{s}, K, \mathcal{M}_e)P(\mathbf{s}, K|\mathcal{M}_t) \tag{2}$$

where $\mathcal{M} = (\mathcal{M}_e, \mathcal{M}_t)$ the set of hyperparameters. The term $p(\mathbf{y}|\mathbf{s}, K, \mathcal{M}_e)$ is the marginal likelihood, conditioned on a partition $\mathbf{s}$ of the data, and is the statistical quantity that all ΔBIC approaches try to maximize. The other term, $P(\mathbf{s}, K|\mathcal{M}_t) = P(\mathbf{s}|K, \mathcal{M}_t)P(K)$ is the prior over partitions $s \in S^n$. Although we did not include it in our ETAPE submission, we will use it in order to penalize partitions with fast changes between speakers. In order to do so, we assume that the overall model is an HMM with gaussian as emission, states correspond to speakers, and $\mathbf{s}$ corresponds to the state sequence. By integrating out the transition probability matrix and the entry probabilities using Dirichlet priors, we end-up with a closed-form expression that places low probability to those partitions with fast changes. The hyperparameters $\mathcal{M}_t = (\alpha_s, \alpha_t)$ represent self-transitions and transitions to other states, by which the Dirichlet priors are parametrized with. The expression of the prior can be found in [12].
For the kth speaker parameters (μ_k, Σ_k), a Normal - Inverse Wishart conjugate prior is attached to each Gaussian, i.e.

$$(\mu_k, \Sigma_k) \sim \mathcal{N}(\mu_0, \frac{1}{\nu}\Sigma_0)\mathcal{IW}(\Psi, p) \tag{3}$$

where $\mathcal{M}_e = (\mu_0, \Psi, \nu, p)$ are the model hyperparameters that denote mean, covariance, number of virtual observations for the Normal prior and degrees of freedom (*dof*) for the Inverse-Wishart prior, respectively. Let $0 < r \leq 1$ be the scaling factor, due to autocorrelation of the features. By integrating out (μ_k, Σ_k) we obtain

$$p(\mathbf{y}_k|\mathcal{M}_e) = \left(\frac{\nu}{rn_k + \nu}\right)^{\frac{d}{2}} \pi^{-\frac{rn_k d}{2}} \frac{|\Psi|^{\frac{p}{2}}}{|\Psi + P_k|^{\frac{rn_k+p}{2}}} \frac{\Gamma_d(\frac{p+rn_k}{2})}{\Gamma_d(\frac{p}{2})} \tag{4}$$

where

$$P_k = r\sum_{i:s^{(i)}=k}(\mathbf{y}^{(i)} - \mu_0)(\mathbf{y}^{(i)} - \mu_0)^T \tag{5}$$

and $\Gamma_d(\cdot)$ the d-variate Gamma function. Finally, the overall marginal likelihood is given by $p(\mathbf{y}|\mathbf{s}, K, \mathcal{M}_e) = \prod_{k=1}^{K} p(\mathbf{y}_k|\mathcal{M}_e)$.

3.1.2. Estimating the partition with Hierarchical Clustering

To maximize $P(\mathbf{s}, K|\mathbf{y}, \mathcal{M})$ with respect to $(\mathbf{s}, K)$, the standard AHC is used. Note that when maximizing, we only consider $K = \max(\mathbf{s})$. However, all events $(\mathbf{s}, K)$ for which $K \geq \max(\mathbf{s})$ have nonzero probability. A partition $\mathbf{s}$ can be the outcome of a Markov model with $K \geq \max(\mathbf{s})$ number of states.
Focusing on maximizing only $p(\mathbf{y}|\mathbf{s}, K, \mathcal{M})$ with AHC, the similarity between $\mathbf{y}_a$ and $\mathbf{y}_b$ is defined by

$$\text{LLR}_{\mathbf{y}_a,\mathbf{y}_b} = \log p(\mathbf{y}_a|\mathcal{M}_e) + \log p(\mathbf{y}_b|\mathcal{M}_e) - \log p(\mathbf{y}_{a\cup b}|\mathcal{M}_e) \quad (6)$$

where each term is defined in eq. (4). Note that this is the closed-form expression of the ΔBIC test discussed above. Moreover, by considering the Laplace approximation to $p(\mathbf{y}_k|\mathcal{M}_e)$, one may verify that it corresponds to the segmental-BIC. Both other settings imply sample-size dependent priors, even when $\lambda = 1$, [12].

3.1.3. Two notes regarding the prior on partitions

To incorporate the partition prior into the AHC algorithm, we should be estimating the probability of s' , where s' the partition after merging a pair of clusters, for every single iteration and pair. Since this procedure can be time consuming, we proceed as follows. For each iteration, the N_p pairs having the smaller LLR (without including the partition prior) are found and their LLRs are stored ($N_p = 5$ is sufficient). These N_p pairs will be the candidates for merging. Then, for the candidate pairs, the difference between the prior log-probability of the current partition (which is common to all pairs, given the iteration count) minus the prior log-probability of s' is calculated and added to the corresponding LLRs. The LLRs - that are now augmented with the partition prior - are sorted again and the first pair in the rank is merged if and only if LLR< 0.
It is interesting to note that the values of the prior would have been negligible compared to those of $p(\mathbf{y}|\mathbf{s}, K, \mathcal{M}_e)$, unless the ESS was used in the latter expression. By using the nominal sample sizes, the likelihood terms tend to dominate completely the posterior of the partition, due to the artificial boost in confidence in the estimates of $\{\mu_k, \Sigma_k\}_{k=1}^{K}$, caused by the use of the nominal sample sizes. And as the experiments show, the role of the partition prior is crucial, because it reduces significantly the number of short segments that are falsely creating new speakers.

3.2. Experimental results

We submitted the algorithm to the ETAPE contest for speaker diarization, as CRIM's secondary system, [13]. The ETAPE is the continuation of the ESTER contests, and is based on broadcasts from several french TV and radio stations. The development (DEV) and evaluation (EVAL) sets consists of 15 shows each, with durations ranging from 10 to 60 minutes. The corpus is divided into two sets, DGA which contains political debates, and ELDA, which contains typical Broadcast News shows.
The tuning of the model parameters was based on the DEV set. This includes the estimation of hyperparameters of the model as well as the tuning of the scaling factor r. The number of virtual observations was set equal for both the Gaussian and the inverse-Wishart prior, $\nu = p = 200r$. The mean of the Gaussian prior was set equal to 0, while Σ_0 equal to the covariance of the DEV set, averaged across shows. Moreover, we set $\Psi = p\Sigma_0$ to respect the conjugacy of the prior to the likelihood. Finally, the scaling factor r was set equal to 0.30 and 0.17 for the DGA and ELDA set, respectively.
We compare the results of the proposed algorithm to those obtain by CRIM's primary system, described in [14]. Like the proposed method, the system uses the AHC algorithm to merge segments. However, it models speakers using a 256-component GMM, based on GMM-UBM adaptation scheme, and uses the normalized cross-likelihood ratio (NCLR) as a similarity measure.
The systems are the same, up to the clustering stage. We should note that the speaker turn detector is using a Viterbi algorithm to locate the boundaries turn more precisely. Due to it, we found that for the proposed algorithm, no further gain could be attained by applying a Viterbi algorithm after the clustering stage. Therefore, the results for the proposed method are simply those obtained by the AHC algorithm. We denote by CRIM$_p$, CRIM$_s$ and CRIM$_s^*$ the GMM-UBM algorithm, the proposed without the partition prior and the proposed with the prior, respectively. The results are given in Table 1 in terms of Diarization Error Rate (DER) and estimated number of speakers. The false alarm rate was 1.4% for 2.0% for DEV and EVAL, respectively for all the systems (since they shared the same speech detector) and the missed detection close to 0%. They clearly demonstrate the strength of the method, considering that a single Gaussian was deployed to model the highly multimodal speaker distribution. Moreover, with the inclusion of the prior on the partition space, we managed to discard the majority of those short segments that formed singleton clusters. This is evident by comparing the estimated number of speakers in CRIM$_s$ and CRIM$_s^*$. Even without the prior, though, the proposed system was ranked amongst the best in the ETAPE contest, when all the other submitted systems (29 on total) used either GMMs or iVectors.

Table 1. *DER (%) and estimated number of speakers (#SPK) on the ETAPE contest. The last column indicates the true number of speakers for each set.*

	CRIM$_p$	CRIM$_s$	CRIM$_s^*$	#SPK
DER(%), DEV	13.48	13.31	13.10	-
#SPK, DEV	173	178	157	152
DER(%), EVAL	19.77	18.08	17.40	-
#SPK, EVAL	184	190	158	156

4. EFFECTIVE SAMPLE SIZE IN IVECTORS

We are now dealing with the problem of speaker recognition. In this section, we show how to apply the effective sample size to the dominant representation of current state-of-the-art technology, the iVectors, [1]. Experimental results will be presented on NIST-2010 8conv-extended, using a standard Gaussian PLDA.

4.1. Short description of iVectors

Recall that iVectors encode the means of a Gaussian Mixture Model (GMM) with typically $C = 2048$ components, which are constrained to lie on a low-dimensional space of $d \in [400, 600]$ dimensions. For the cth mixture component, let $\mathbf{x} \in \Re^d$ be an iVector, $(w_c, m_c, \Sigma_c)_{c=1}^{C}$ the parameters of the Universal Background Model (UBM) and V_c the cth block of the total variability matrix $V^T = [V_1^T, \ldots, V_C^T]$. Then, the mean vector of the cth

component μ_c is given by the following equation

$$\mu_c = m_c + V_c \mathbf{x} \tag{7}$$

In order to extract the iVector of an utterance $\mathbf{y} = \{\mathbf{y}^{(i)}\}_{i=1}^{n}$, the zeros and first order Baum-Welch statistics are calculated as follows

$$(N_c, F_c) = \left(\sum_{i=1}^{n} \gamma_c^{(i)}, \sum_{i=1}^{n} \gamma_c^{(i)} \mathbf{y}^{(i)} \right) \tag{8}$$

where $\gamma_c^{(i)}$ is the posterior of the ith frame to belong to the cth mixture component. Then, the iVector $\tilde{\mathbf{x}}$ (i.e. the posterior expectation or MAP estimate of $\mathbf{x}$) is given by the following formula

$$\tilde{\mathbf{x}} = \mathrm{Cov}(\mathbf{x}) \sum_c V_c^T \Sigma_c^{-1} (F_c - N_c m_c) \tag{9}$$

where

$$\mathrm{Cov}(\mathbf{x}) = \left(I + \sum_c N_c V_c^T \Sigma_c^{-1} V_c \right)^{-1} \tag{10}$$

the covariance of the estimate of $\mathbf{x}$, [1].

4.2. Scaling-down the statistics

Scaling the Baum-Welch statistics is straightforward. We simply apply $N_c^r = rN_c$ and $F_c^r = rF_c$. Note that the same operation should be applied when training the iVector extractor.
The effect of this scaling is two-fold. First, the volume of $\mathrm{Cov}(\mathbf{x})$ increases in order to reflect the larger uncertainty regarding the estimate, while the second effect would be the increased contribution of the prior on $\tilde{\mathbf{x}}$. Although the results we present do not make use of the uncertainty in the estimate, they demonstrate how the second effect alone can improve the recognition performance.

4.3. Experiments on NIST-2010

We performed experiments on the *8conv - coreext* condition of the telephone speech NIST extended list. We focus on female data only, where the state-of-the-art performance is worst than the one on male data. We use the Equal Error Rate (EER) and the (new and old) minimum Detection Cost Function (minDCF) of NIST as metrics. For specifications regarding the UBM and iVector extractor we refer to [15].
The model we use is the standard Gaussian PLDA with length normalization, [16]. No averaging is applied to neither the iVector, nor to the LLR level. We have also formed a second experiment, where the number of enrollment recordings was randomized between 1 and 8. The results are given in Table 2 (denoted by 8conv), while the DET curves are illustrated in Fig. 1.

Table 2. *Results on NIST-2010 8conv and 1-8conv extended female telephone data (i.e. det5)*

	EER(%)	minDCFold	minDCFnew
8conv, r = 1	1.26	0.065	0.28
8conv, r = 1/3	1.22	0.062	0.24
1-8conv, r = 1	2.25	0.080	0.30
1-8conv, r = 1/3	2.14	0.072	0.28

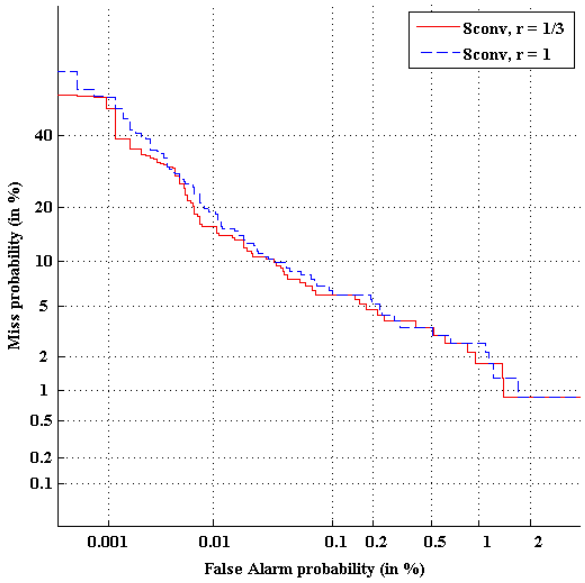

Fig. 1. *DET curves on NIST-2010 8conv extended female telephone data (i.e. det5). Blue dashed line: no scaling factor (i.e. $r = 1$), Red line: scaling factor $r = 1/3$.*

The results show that an improvement is attained in all metrics for both 8conv and 1-8conv sets. The interpretation is that when the scaling is omitted, the posterior expectation of $\mathbf{x}$ is dominated by the likelihood and therefore becomes too noisy. By using a reasonable scaling factor, the prior manages to regularize the iVectors, leading the PLDA model to increased performance. We should also emphasize that this improvement has been attained without full optimization of the scaling factor r. The only other value we examined was $r = 1/5$, which resulted in inferior performance compared to $r = 1/3$, yet slightly better to the one with $r = 1$.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we showed the reason why the statistics of features need to be scaled down, and presented the benefits from such a normalization. We focused on two major tasks in speaker characterization technologies, namely diarization and recognition, and on models of highly diverge complexity. The experiments on diarization were based on the ETAPE contest and showed how competitive a baseline approach can be, when the ESS is taken into account. Moreover, the experiments on iVector-based speaker recognition showed an improvement in performance, ranging from small to significant, depending on the score metric.
For future work, our priority will be to examine whether the automatic extraction of the scaling factor proposed in [7] works well in the case of speech. If so, several possibilities would open, such as the use of utterance-dependent scaling factors. This scenario seems to be reasonable, since inter-frame correlation varies across speakers and even speech segments of the same speaker. Moreover, we are planning to examine the effectiveness of the ESS scaling in methods that make use of the uncertainly of the iVector in the back-end, [3]. Finally, it would be interesting to reconsider several algorithms proposed in diarization and recognition literature, using the ESS scaling. The state-of-the-art performance of the AHC algorithm showed that it can be accounted as a good example.

6. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", in *IEEE Transactions on Audio, Speech & Language Processing*, 2011.

[2] J. Villalba and N. Brummer, "Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance", *Proc. of Interspeech 2011*, Florence, Italy.

[3] P. Kenny. T. Stafylakis, P. Ouellet, M.J. Alam and P. Dumouchel, "PLDA for Speaker Verification with Utterances of Arbitrary Duration", submitted to ICASSP '13.

[4] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky, The Sticky HDP-HMM: Bayesian nonparametric Hidden Markov Models with Persistent States, 2009.

[5] S. Shum, N. Dehak, and J. Glass, "On the Use of Spectral and Iterative Methods for Speaker Diarization," in Proceedings of Interspeech, Portland, Oregon, 2012.

[6] S. Yue, P. Pilon, B. Phinney and G. Cavadias. "The influence of autocorrelation on the ability to detect trend in hydrological series", in Hydrological Processes, Vol. 16, 2002.

[7] D. P. Lettenmaier, "Detection of Trends in Water Quality Data From Records With Dependent Observations", Journal of Water Resources Research, Vol. 2, No. 5, 1976.

[8] G. Schwarz, Estimating the dimension of a model, Annals of Statistics, vol. 6, 1978.

[9] S.S. Chen and P.S. Gopalakrishnam, Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion, in Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[10] X. Zhu, C. Barras, S. Meignier, and J. Gauvain, Combining Speaker Identification and BIC for Speaker Diarization, in Proceedings of Interspeech, 2005.

[11] T. Stafylakis, V. Katsouros, and G. Carayannis, The Segmental Bayesian Information Criterion and its applications to Speaker Diarization, IEEE Selected topics in Signal Processing, October 2010.

[12] T Stafylakis, X Anguera, V Katsouros, G Carayannis, "Closed-form expressions vs. BIC: A comparison for speaker clustering", in proc. of Acoustics, Speech and Signal Processing (ICASSP), 2011.

[13] G. Gravier, G. Adda, N. Paulsson, M. Carre, A. Giraudel, O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language", Technical Report, 2011.

[14] V. Gupta, G. Boulianne, P. Kenny, P. Ouellet, P. Dumouchel, Speaker Diarization of French Broadcast News, ICASSP-2008.

[15] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, P. Dumouchel, "Mixture of PLDA models in i-vector space for gender independent speaker recognition", in *Proc. of Interspeech 2011*, Florence, Italy.

[16] D. Garcia-Romero and C. Y. Espy-Wilso, Analysis of i-vector length normalization in speaker recognition systems, in *Proceedings of Interspeech*, Florence, Italy, Aug. 2011.