UNSUPERVISED ADAPTATION OF PLDA BY USING VARIATIONAL BAYES METHODS

Jesús Villalba, Eduardo Lleida

VIVOlab, Aragon Institute for Engineering Research (I3A),
University of Zaragoza, Spain
{villalba,lleida}@unizar.es

ABSTRACT

State-of-the-art speaker recognition relays on models that need a large amount of training data. This models are successful in tasks like NIST SRE because there is sufficient data available. However, in real applications, we usually do not have so much data and, in many cases, the speaker labels are unknown. We present a method to adapt a PLDA model from a domain with a large amount of labeled data to another with unlabeled data. We describe a generative model that produces both sets of data where the unknown labels are modeled like latent variables. We used variational Bayes to estimate the hidden variables. We performed experiments adapting a model trained on Switchboard to NIST SRE without labels. The adapted model is evaluated on NIST SRE10. Compared to the non-adapted model, EER improved by 42% and 49% by adapting with 200 and with all the NIST speakers respectively.

Index Terms— speaker recognition, PLDA, i-vector, unsupervised adaptation, variational Bayes

1. INTRODUCTION

The i-vector approach is the state-of-the-art for speaker verification. It provides a method to map a speech utterance to a low dimensional fixed length vector retaining the speaker identity [1]. Great performance has been achieved by modeling the i-vectors distributions by a generative model known as PLDA [2–4]. PLDA needs to be trained on databases with a large number of speakers and sessions. In NIST evaluations [5], enough data is available, however, in many real applications the amount of development data is limited and, in many cases, the speaker labels are unknown.

There are previous works that address the problem of database mismatch with PLDA models. In [6], dataset shift was prevented by normalizing each i-vector by its magnitude (length normalization). Thus, development and test i-vector distributions are made closer. In [7], several adaptation techniques were applied to mitigate language mismatch being length normalization the one attaining better results.

In [8, 9], authors computed fully Bayesian likelihood ratios by integrating out the parameters of the PLDA model. This helps with dataset shift, because the posterior distributions that result, if the amount of training data is small, are heavy-tailed.

In [10], we presented a variational Bayes (VB) method to adapt a full-rank PLDA model from one domain to another with scarce development data. In this paper, we continue that work with a new difficulty added: the labels of the adaptation data are unknown. To test our method we adapted a model trained on Switchboard to NIST SRE. This task was proposed in the recent JHU workshop on speaker recognition [1].

2. UNSUPERVISED SPLDA

2.1. Model description

Simplified probabilistic linear discriminant analysis (SPLDA) is a linear generative model that assumes that an i-vector ϕ_j of speaker i can be written as:

$$\phi_j = \mu + \mathbf{V}\mathbf{y}_i + \epsilon_j \tag{1}$$

where μ is a speaker independent term, $\mathbf{V}$ is a low rank eigenvoices matrix, $\mathbf{y}_i$ is the speaker factor vector, and ϵ_j is the within class variability term. We put a standard normal prior on $\mathbf{y}_i$ and normal with zero mean and precision $\mathbf{W}$ on ϵ_j.

Figure 1 depicts the Bayesian network of this model where the training data has been split into two parts: one with known labels (out-of-domain data), and another with hidden labels (in-domain or adaptation data). We denote the out-of-domain data by the subscript d.

θ_d are the labels of the out-of-domain data and partition N_d i-vectors into M_d speakers. θ are the labels of the in-domain data and partition N i-vectors into M speakers. θ_j is a latent variable comprising a 1–of–M binary vector with elements θ_{ji} with $i = 1, \ldots, M$. Note that, the distribution of each speaker is assumed to be Gaussian with mean $\mu + \mathbf{V}\mathbf{y}_i$ and precision $\mathbf{W}$. The set of all the speakers form a GMM

Thanks to the JHU CSLP for hosting the workshop where this work started, all the workshop participants for fruitful discussion and Doug Reynolds for setting up the adaptation task. This work has been supported by the Spanish Government and the European Union (FEDER) through projects TIN2011-28169-C05-02 and INNPACTO IPT-2011-1696-390000.

[1] http://www.clsp.jhu.edu/workshops/archive/ws13-summer-workshop/groups/spk-13/

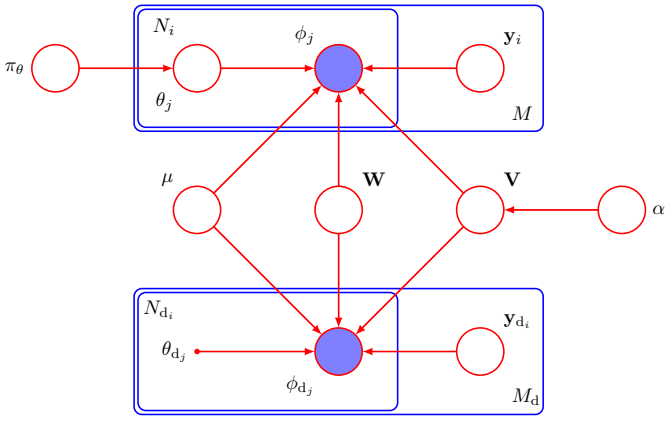

Fig. 1. BN for unsupervised SPLDA.

where θ correspond to the component occupations. The conditional distribution of θ given the mixture weights π_θ is

$$P(\theta|\pi_\theta) = \prod_{j=1}^{N}\prod_{i=1}^{M} \pi_{\theta_i}^{\theta_{ji}} \ . \tag{2}$$

We put a Dirichlet prior on the weights:

$$P(\pi_\theta|\tau_0) = \mathrm{Dir}(\pi_\theta|\tau_0) = C(\tau_0)\prod_{i=1}^{M} \pi_{\theta_i}^{\tau_0-1} \tag{3}$$

where by symmetry we choose the same τ_0 for all the components, and $C(\tau_0)$ is the normalization constant.

2.2. Model priors

We chose the model priors based on Bishop's paper about VB PPCA [11]. We introduced a *hierarchical* prior $P(\mathbf{V}|\alpha)$ over $\mathbf{V}$ through a conditional Gaussian distribution of the form:

$$P(\mathbf{V}|\alpha) = \prod_{q=1}^{n_y}\left(\frac{\alpha_q}{2\pi}\right)^{d/2} \exp\left(-\frac{1}{2}\alpha_q \mathbf{v}_q^T\mathbf{v}_q\right) \tag{4}$$

where $\mathbf{v}_q$ are the columns of $\mathbf{V}$ and n_y is the speaker factors dimension. Each α_q controls the inverse variance of the corresponding $\mathbf{v}_q$. If a particular α_q has a posterior distribution concentrated at large values, the corresponding $\mathbf{v}_q$ will tend to be small, and that direction of the latent space will be effectively 'switched off'.

We defined a prior for α:

$$P(\alpha) = \prod_{q=1}^{n_y} \mathcal{G}(\alpha_q|a_\alpha, b_\alpha) \tag{5}$$

where $\mathcal{G}$ denotes the Gamma distribution.

We placed a Gaussian prior for the mean μ:

$$P(\mu) = \mathcal{N}\left(\mu|\mu_0, \beta^{-1}\mathbf{I}\right) \ . \tag{6}$$

Finally, we put Wishart priors on $\mathbf{W}$ that can be non-informative (Jeffreys prior) like

$$P(\mathbf{W}) = \lim_{k\to 0} \mathcal{W}(\mathbf{W}|\mathbf{W}_0/k, k) \quad = \alpha\,|\mathbf{W}|^{-(d+1)/2} \tag{7}$$

or informative like

$$P(\mathbf{W}) = \mathcal{W}(\mathbf{W}|\mathbf{\Psi}_0, \nu_0) \ . \tag{8}$$

2.3. Variational Bayes with deterministic annealing

We approximated the joint posterior of the latent variables by a factorized distribution of the form:

$$P(\mathbf{Y}, \mathbf{Y}_{\mathrm{d}}, \theta, \pi_\theta, \mu, \mathbf{V}, \mathbf{W}, \alpha|\mathbf{\Phi}, \mathbf{\Phi}_{\mathrm{d}}) \approx$$
$$q(\mathbf{Y}, \mathbf{Y}_{\mathrm{d}})\, q(\theta)\, q(\pi_\theta) \prod_{r=1}^{d} q(\tilde{\mathbf{v}}'_r)\, q(\mathbf{W})\, q(\alpha) \tag{9}$$

where $\tilde{\mathbf{v}}'_r$ is a column vector containing the r^{th} row of $\tilde{\mathbf{V}} = [\mathbf{V} \quad \mu]$. If $\mathbf{W}$ were diagonal the factorization $\prod_{r=1}^{d} q(\tilde{\mathbf{v}}'_r)$ would not be necessary because it would arise naturally. However, for full $\mathbf{W}$, we have to force the factorization to make the problem tractable.

We computed these factors by using Variational Bayes [12] with deterministic annealing (DA) [13]. The formula to update a factor q_j is

$$\ln q_j^*(\mathbf{Z}_j) = \mathrm{E}_{i\neq j}\left[\kappa \ln P(\mathbf{\Phi}, \mathbf{\Phi}_{\mathrm{d}}, \mathbf{Z})\right] + \mathrm{const} \tag{10}$$

where $\mathbf{Z}$ abbreviates the set of all hidden variables, $\mathbf{Z}_j$ are the hidden variables corresponding to the j^{th} factor, and κ is the annealing factor; expectations are taken with respect to all the factors $i \neq j$. Equation (10) optimizes the VB lower bound

$$\mathcal{L} = \mathrm{E}\left[\ln P(\mathbf{\Phi}, \mathbf{\Phi}_{\mathrm{d}}, \mathbf{Z})\right] - \mathrm{E}\left[\ln q(\mathbf{Z})\right] \tag{11}$$

where expectations are taken with respect to the variational posterior $q(\mathbf{Z})$. $\mathcal{L}$ approximates $\ln P(\mathbf{\Phi}, \mathbf{\Phi}_{\mathrm{d}})$. Annealing modifies the VB objective in a way that helps to avoid local maxima. We must set $\kappa < 1$ at the start and increase it in each iteration until $\kappa = 1$. The terms of $\ln P(\mathbf{\Phi}, \mathbf{\Phi}_{\mathrm{d}}, \mathbf{Z})$ corresponding to the out-of-domain data were scaled by another parameter $\eta < 1$ to reduce its weight on the model posterior.

The full VB equations can be found in our report [14].

3. EXPERIMENTAL SETUP

3.1. Task description

We tested our method by adapting a SPLDA trained on Switchboard (SWB) (out-of-domain) to NIST SRE (in-domain). This task was proposed during the last JHU workshop on speaker recognition. The out-of-domain data consisted of 33068 segments from 3114 speakers with known labels. The in-domain data consisted of 36706 segments from 3807 speakers of NIST SRE04-08 with unknown labels. To perform faster experiments, we also created random subsets of 200 and 500 in-domain speakers. The adapted models were evaluated on the NIST SRE10 det5 (tel-tel) extended condition.

3.2. i-vectors

The JHU HLT-COE provided 600 dimensional i-vectors for this work. They were extracted using 20 MFCC + Δ with short time mean and variance normalization. The UBM and i-vector extractor were gender independent and used 2048 Gaussians. We applied centering, whitening and length normalization to the i-vectors [6]. The parameters needed for centering and whitening were trained from all the NIST SRE data since, for that, speaker labels are not required.

3.3. SPLDA

Our SPLDA models were gender independent with speaker factors of dimension 150. We tried two types of priors for the parameters of the SPLDA: non-informative and informative. For the non-informative case, we chose wide priors for μ and α by setting $\mu_0 = \mathbf{0}$ and $\beta = a_\alpha = b_\alpha = 10^{-3}$; and Jeffreys prior for $\mathbf{W}$.

For the informative case, we chose our priors based on the average total variance of the data s_0^2 (average across dimensions). We observed that, for a SPLDA trained on SWB with i-vectors centered and whitened with parameters also trained on SWB, the average variance of the speaker space was approximately 15% of s_0^2 and the channel variance was the remaining 85%. To set our priors, we assumed that, for NIST SRE, those percentages could be similar . Thus, we computed s_0^2 from the adaptation data. Then, for α (prior of the inverse eigenvalues), we placed a wide prior with mode $1/(0.15s_0^2)$ by setting $a_\alpha = 2$ and $b_\alpha = 0.15s_0^2$. For $\mathbf{W}$, we used a Wishart prior with expectation $1/(0.85s_0^2)\mathbf{I}$ by setting $\nu_0 = 602$ and $\Psi_0 = 1/(0.85s_0^2\nu_0)\mathbf{I}$. Note that, for the Wishart prior to be proper, we need $\nu_0 > d$, this means that the prior will have an important influence on the posterior unless that we have a number of training segments $N >> d$.

The expectations of the model parameters given the VB posteriors were used to compute the likelihood ratios of the evaluation set in the standard way.

We also tried a simplified model without priors on the SPLDA parameters. In this case, the model parameters are point estimates computed by maximizing $\mathcal{L}$.

We set the parameter η that controls the weight of the out-of-domain data to 0.25. This value produced the lowest error rate when training the model on labeled in-domain data. How to select this parameter in an unsupervised way remains an open problem that we do not treat in this work.

3.4. Model selection

To select the optimum number of speakers of the unlabeled dataset, we tried to initialize the algorithm assuming a large number of speakers and, iteration by iteration, eliminate the speakers with smaller number of samples. This method is similar to the *automatic relevance determination* (ARD) explained in [12] to find the number of components in a GMM.

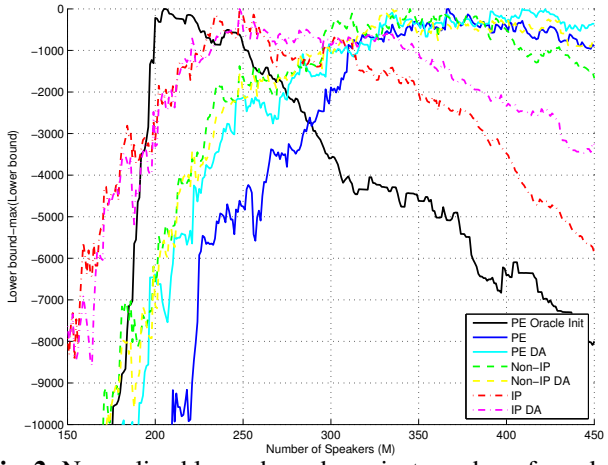

Fig. 2. Normalized lower bounds against number of speakers (M) for the case where the actual M=200.

We tried several criteria to prune speakers. However we were only able to merge a small number of speakers.

Then, we applied a brute force approach where we ran the algorithm several times, each time hypothesizing a different number of speakers M. We selected the best model based on the VB lower bound $\mathcal{L}(M)$. To rightfully compare lower bounds for different number of speakers, we need to set the parameter of the Dirichlet prior on the speakers weights to $\tau_0 = 400/M$. To select the value 400, we tried several values and chose the one that produced the largest sum of lower bounds $\sum_M \mathcal{L}(M)$.

In the experiments with annealing, we initialized the annealing parameter to $\kappa = 0.3$ and, in each iteration, we updated κ as $\kappa \leftarrow 1.1\kappa$. We tried several initial values for κ and chose the one that provided the largest final lower bound.

3.5. Initialization

The SPLDA was initialized from the model trained on SWB. To initialize the speaker labels, we computed the matrix of likelihood ratios of all against all in-domain i-vectors with the initial model. Then, we chose the proper threshold to partition the data into the desired number of speakers.

We also did a cheating experiment with the training list of 200 speakers. For $M = 200$, we initialized with the true labels; for $M < 200$, we randomly merged speakers; and for $M > 200$, we randomly split them. For example, for $M = 300$, we have 100 speakers with perfect labels and 100 speakers split into two speakers. We denote this experiment by *Oracle Init*.

3.6. EXPERIMENTS RESULTS

Figure 2 compares lower bounds against the number of hypothesized speakers for different adaptation methods and the adaptation list of 200 speakers. For better visualization, the lower bounds are normalized by subtracting the maximum of each method. The methods denoted by PE compute point estimates of the model parameters instead of posterior distribu-

Table 1. *EER(%)/MinDCF for different adaptation options. Table blocks correspond to adapting with 200, 500 and all the speakers.*

	M Actual			M Max $\mathcal{L}$		
	EER(%)	MinDCF	M	EER(%)	MinDCF	M
SWB no adapt	6.57	0.66	-	6.57	0.66	-
PLDA(**I**,**I**)	5.96	0.66	-	5.96	0.66	-
Oracle	**3.27**	**0.52**	200	**3.27**	**0.52**	200
VB PE Oracle Init	3.23	0.52	200	3.24	0.53	205
VB PE	3.87	0.56	200	5.40	0.62	367
VB PE DA	3.64	**0.55**	200	5.85	0.65	409
VB Non-IP	3.91	0.56	200	5.09	0.61	345
VB Non-IP DA	3.73	0.56	200	4.98	0.61	335
VB IP	3.71	**0.55**	200	3.86	**0.55**	248
VB IP DA	**3.61**	**0.55**	200	**3.83**	**0.55**	248
Oracle	**3.02**	**0.50**	500	**3.02**	**0.50**	500
VB PE	3.81	0.56	500	5.04	0.63	868
VB PE DA	3.74	0.55	500	6.21	0.66	1079
VB Non-IP	3.71	0.55	500	5.03	0.63	875
VB Non-IP DA	**3.52**	0.56	500	5.69	0.65	1073
VB IP	3.70	0.55	500	4.21	0.58	726
VB IP DA	3.57	**0.54**	500	**4.01**	**0.57**	691
Oracle	**2.19**	**0.42**	3807	**2.19**	**0.42**	3807
VB PE	3.17	0.54	3807	4.12	0.58	6558
VB PE DA	2.99	0.52	3807	3.66	0.57	7438
VB Non-IP	3.26	0.54	3807	4.04	0.58	6489
VB Non-IP DA	**2.98**	0.52	3807	3.69	0.57	7601
VB IP	3.29	0.53	3807	3.98	0.58	6492
VB IP DA	3.01	**0.51**	3807	**3.37**	**0.56**	6867

tions. The label DA means deterministic annealing, IP means informative priors and Non-IP means non-informative priors.

The oracle initialization has the maximum at $M = 205$ and decays rapidly at both sides. Initializing the labels with the SWB model, we obtain maxima in much higher M values. With the oracle initalization we obtain lower error rate so, ideally, the oracle initialization should provide better lower bound but that did not happen ($\mathcal{L}_{\text{Oracle}} - \mathcal{L}_{\text{SWB}} = -2643$). Thus, we cannot use $\mathcal{L}$ to choose the best initialization.

Regarding the detection of the number of speakers, there is not a significant difference between the model with point estimates or the full model with non-informative priors. The model with informative priors gets the best estimation of the number of speakers.

Table 1 compares EER and minimum DCF for multiple variants of the algorithm and adaptation lists. It also compares the results between choosing the correct number of speakers or the number of speakers with maximum $\mathcal{L}$. In the first two lines, we see that full-rank PLDA with identity between and within covariances (equivalent to cosine distance scoring) achieved better EER than the model trained on SWB. So, in this case, a generic model was better than the model trained on out-of-domain data.

We noted that, if we choose the model corresponding to the actual number of speakers, there are small differences between computing point estimates or posterior distributions. We also noted that the models with DA reach lower error rates. Another consideration to make is how the amount of adaptation data affects the results. By using informative priors and DA, and adapting with 200 speakers, the EER improved by 45% compared to the SWB model. However, by adapting with 500 and 3807 speakers, it only improved by 1% and 16% compared to 200 speakers. As we increase the amount of adaptation data the gap between the results with Oracle and unknown labels grows. When adapting with 200 speakers the unsupervised model is 10% worse than the oracle and, with all the data, it is 37% worse. This seems to indicate that we will reach a point where increasing the amount of unsupervised data will not help anymore.

Now, we look at the column where we choose the model with largest VB lower bound. Models with DA, in most cases, selected a larger number of speakers. This was harmful when adapting with 200 and 500 speakers, except for models with informative priors. When adapting with all the speakers, models with DA, even selecting a larger speaker number, provided the lowest error rates. For all adaptation lists, the best model combined informative priors and DA. Again, the improvement that we obtain as we increase the data becomes smaller. For example, the performance for 500 speakers was worse than for 200. Besides, the minDCF did not improve by adapting with all the speakers with regard to 200 speakers.

4. CONCLUSIONS

We presented a method to adapt a SPLDA model from a domain where we have a large amount of labeled training data to another domain where the speaker labels of the training data are unknown. For that, we designed a generative model that generates both sets of data (labeled and unlabeled) and the unknown labels were modeled as latent variables. We tried three variants of the model distinguished according to how we define the SPLDA parameters: deterministic parameters obtained by maximum likelihood or, latent variables with non-informative or informative priors.

We used a variational Bayes procedure to approximate the posterior distributions of the latent variables involved. Deterministic annealing was used to avoid being trapped in local maxima of the VB lower bound. To detect the number of speakers M of the unlabeled dataset, we ran simulations hypothesizing different values of M and chose the model that obtained the largest VB lower bound. This method detected more speakers than they actually are, however, the model selected in this way was still able to provide good recognition rates. The best results were achieved by combining informative priors and annealing.

We experimented adapting a model from Switchboard to the unlabeled NIST SRE04-08 dataset. We evaluated the adapted model on the NIST SRE10 det5 condition. Compared to the model trained on Switchboard, EER improved by 42% and 49% by adapting with 200 speakers and with all the available data respectively. The minDCF improved by around 15% for all the adaptation lists evaluated. It did not benefit from using a large amount of adaptation data.

5. REFERENCES

[1] Najim Dehak, Patrick Kenny, Redah Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis For Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

[2] Simon J. D. Prince and James H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *Proceedings of the IEEE International Conference on Computer Vision, ICCV 2007*, Rio de Janeiro, Brazil, Oct. 2007, IEEE.

[3] Patrick Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Proceedings of Odyssey 2010 - The Speaker and Language Recognition Workshop*, Brno, Czech Republic, July 2010, ISCA.

[4] Jesús Villalba, Eduardo Lleida, Alfonso Ortega, and Antonio Miguel, "The I3A Speaker Recognition System for NIST SRE12: Post-evaluation Analysis," in *14th Annual Conference of the International Speech Communication Association, Interspeech 2013*, Lyon, France, Aug. 2013, ISCA.

[5] NIST Speech Group, "NIST Speaker Recognition Evaluation," .

[6] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech 2011*, Florence, Italy, Aug. 2011, pp. 249–252, ISCA.

[7] Carlos Vaquero, "Dataset Shift in PLDA based Speaker Verification," in *Proceedings of Odyssey 2012 - The Speaker and Language Recognition Workshop*, Singapore, June 2012, pp. 39–46, COLIPS.

[8] Jesús Villalba and Niko Brummer, "Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech 2011*, Florence, Italy, Aug. 2011, pp. 28–31, ISCA.

[9] Jesús Villalba, Niko Brummer, and Eduardo Lleida, "Fully Bayesian Likelihood Ratios vs i-vector Length Normalization in Speaker Recognition Systems," in *NIST SRE11 Speaker Recognition Workshop*, Atlanta, Georgia, USA, Dec. 2011.

[10] Jesús Villalba and Eduardo Lleida, "Bayesian Adaptation of PLDA Based Speaker Recognition to Domains with Scarce Development Data," in *Proceedings of Odyssey 2012 - The Speaker and Language Recognition Workshop*, Singapore, June 2012, COLIPS.

[11] Christopher Bishop, "Variational principal components," in *Proceedings of the 9th International Conference on Artificial Neural Networks, ICANN 99*, Edinburgh, Scotland, Sept. 1999, IET, pp. 509–514.

[12] Christopher Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, 2006.

[13] Kentaro Katahira, Kazuho Watanabe, and Masato Okada, "Deterministic annealing variant of variational Bayes method," *Journal of Physics: Conference Series International Workshop on Statistical-Mechanical Informatics 2007 (IW-SMI 2007)*, vol. 95, Jan. 2008.

[14] Jesús Villalba, "Unsupervised Adaptation of SPLDA," Tech. Rep., University of Zaragoza, Zaragoza (Spain), 2013.