TOWARDS PLDA-RBM BASED SPEAKER RECOGNITION IN MOBILE ENVIRONMENT: DESIGNING STACKED/DEEP PLDA-RBM SYSTEMS

Andreas Nautsch[⋆] *Hong Hao*[⋆†] *Themos Stafylakis*[‡] *Christian Rathgeb*[⋆] *Christoph Busch*[⋆]

[⋆]da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
{andreas.nautsch,christian.rathgeb,christoph.busch}@{cased|h-da}.de
[†]Department of Applied Mathematics and Computer Science, Technical University of Denmark
s130616@student.dtu.dk
[‡]Centre de Recherche Informatique de Montréal (CRIM), Canada
themos.stafylakis@crim.ca

ABSTRACT

The vast majority of text-independent speaker recognition systems rely on intermediate-sized vectors (i-vectors), which are compared by probabilistic linear discriminant analysis (PLDA). This paper proposes a PLDA-alike approach with restricted Boltzmann machines for i-vector based speaker recognition: two deep architectures are presented and examined, which aim at suppressing channel effects and recovering speaker-discriminative information on back-ends trained on a small dataset. Experiments are carried out on the MOBIO SRE'13 database, which is a challenging and publicly available dataset for mobile speaker recognition with limited amounts of training data. The experiments show that the proposed system outperforms the baseline i-vector/PLDA approach by relative gains of 31% on female and 9% on male speakers in terms of half total error rate.

Index Terms— PLDA-RBM, deep learning, speaker recognition, MOBIO

1. INTRODUCTION

In past years, speaker recognition systems based on intermediate-sized vectors (i-vectors) [1] became state-of-the-art biometric features in combination with a Gaussian probabilistic linear discriminant analysis (G-PLDA) [2] back-end. Recently [3], non-linear PLDA schemes have been shown to be applicable on the domain shift JHU-2013 i-vector set [4], where further gains were also yielded by deep architectures.

In 2012 a proof-of-concept [5] illustrated, that restricted Boltzmann machines (RBMs) can be designed, such that the behaviour of conventional PLDA is achieved at comparable performance to PLDA (PLDA-RBM). Similarly to PLDA, PLDA-RBM extracts latent speaker factors by removing channel effects, i.e.: hidden speaker and hidden channel units. Conceptually, i-vectors are reconstructed by the activation of hidden speaker units, which are then used for comparison. While RBMs in machine learning usually rely on non-linear energy functions, [5] presumed linear energy functions following the conventional Gaussian assumption. Our work examines the impact of Bernoulli energy functions, introduces two stacking approaches, and analyses the information reinforced by each layer. Aiming at mobile environments, experiments are carried out on the publicly available MOBIO speaker recognition evaluation task (MOBIO SRE'13) [6, 7], which *provides a challenging and realistic test-bed for current state-of-the-art speaker verification* [7]. Two concepts for deep PLDA-RBM are examined for compensating channel effects still being persistent on i-vector back-ends trained on limited data sets. The proposed concept is shown to be applicable for systems operating on limited mobile data.

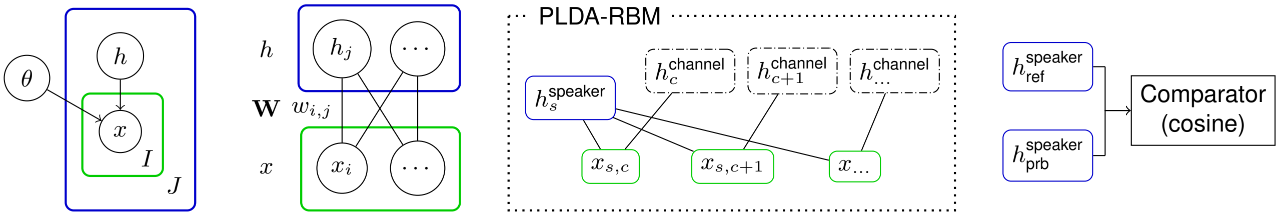
This paper is organized as follows: Section 2 introduces related work, Section 3 describes PLDA-RBMs and the proposed deep designs, Section 4 presents and discusses experimental analyses results, and Section 5 concludes.

2. RELATED WORK

2.1. PLDA-based Speaker Recognition

State-of-the-Art speaker recognition techniques rely on generative pairwise models [8]. In PLDA, an i-vector $i_{s,c}$ is decomposed into a speaker- and channel-independent mean μ, latent speaker factors y_s, and residual noise ϵ_c, where s, c denote speaker- and channel-dependencies: $i_{s,c} = \mu + \mathbf{\Phi}\, y_s + \epsilon_c$ with $y_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\epsilon_c \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$, where $\mathbf{\Phi}$ constrains the low-rank speaker factor subspace, and $\mathbf{\Lambda}$ is the covariance matrix of the residual noise ϵ_c. A speaker verification score S is computed as a log-likelihood ratio of the hypotheses, a reference and a probe i-vector $i_{\text{ref}}, i_{\text{prb}}$ are (a) belonging to the same speaker, or (b) not, which is analytically evaluated by marginal Gaussian likelihoods [9].

978-1-4799-9988-0/16/$31.00 ©2016 IEEE



(a) PLDA [10] (b) RBM [11] (c) PLDA-RBM enrolment [5] (d) PLDA-RBM verification [5]

Fig. 1. Factorization concepts, with: data samples x, hidden variables h, PLDA parameters $\theta = \{\mu, \mathbf{\Phi}, \mathbf{\Lambda}\}$, weights $\mathbf{W}$.

2.2. Restricted Boltzmann Machines

A RBM is a bipartite undirected graphical model, with no connections between units of the same layer [12]. This property makes the distributions of the two layers conditionally independent, and hence allows fast sampling-based training techniques to apply. RBMs can serve different purposes e.g., probabilistic principal component analysis (PPCA), feature reconstruction, or unsupervised initialization of Deep Neural Networks (DNNs) [5, 11, 13, 14].

RBMs are two-layer structure models containing a visible and a hidden layer $v = \{v_i\}_{i=1,\dots,d_v}, h = \{h_j\}_{j=1,\dots,d_h}$, which are connected through a weight matrix $\mathbf{W} = \{w_{i,j}\}$ [5, 11]. The joint probability density function (pdf) of (v, h) is a $(d_v + d_h)$-dimensional Gaussian depending on an energy function $E(v, h)$: $P(v, h \,|\, \mathbf{W}) = Z^{-1} \exp(\, -E(v, h)\,)$ with normalizing constant Z.

Energy functions model the distribution of visible and hidden units: *Gaussian-Gaussian* (GG) layers assume Gaussian distribution for visible and hidden units, *Gaussian-Bernoulli* (GB) layers assume Gaussian distribution for visible units and Bernoulli distribution for hidden units. By assuming zero mean for GG energy functions, the distributions take the form of PPCA [5, 13]. The GB energy function $E(v, h)$ considers the hidden unit pdf to be Bernoulli-distributed [13, 15], such that the GB energy function takes the form of [13]:

$$E(v,h) = \sum_{i\in d_v} \frac{v_i^2}{2\,\sigma_i^2} - \sum_{j\in d_h} b_j\, h_j - \sum_{\substack{i\in d_v\\ j\in d_h}} \frac{v_i}{\sigma_i}\, h_j\, w_{ij}. \qquad (1)$$

3. PLDA-RBM AND DEEP PLDA-RBM

RBMs can be used in a specific way to achieve a PLDA-similar back-end, i.e. visible units representing i-vectors can be decomposed into hidden speaker units h_s^{speaker} and hidden channel units h_c^{channel} representing speaker and channel/residual factors, respectively. Fig. 1c depicts the basic idea: during enrolment, speaker-dependent RBM weights $\mathbf{W}(s)$ are learned constrained to purify h_s^{speaker}. During verification, the same weights $\mathbf{W}(s)$ are used for purifying probe i-vectors [5].

3.1. Basic PLDA-RBM Algorithm

We are following a similar approach to [5]. Our main difference is the usage of Bernoulli hidden layers, i.e. a GB PLDA-RBM. The PLDA-RBM is trained with CD1 using mini-batches and standard L2 regularization, while no momentum terms are added [13]. During recognition phase, features are extracted using the speaker layer, one per i-vector. In the case of multi-sample enrolments, references are created as averaged template features. Reference and probe features are compared by cosine distance. Fig. 1 depicts PLDA, RBM, and PLDA-RBM architectures.

3.2. Deep Designs

One of the motivation behind the i-vector paradigm was the insufficiency of JFA in distinguishing between speaker and channel information, as channel factors were shown to containing speaker information [1]. A two-step approach, where a total variability subspace is first estimated, followed by a back-end (PLDA, LDA with cosine distance, a.o.) that distinguishes the two types of variability on the i-vector space proved to be superior to the JFA monolithic classifier.

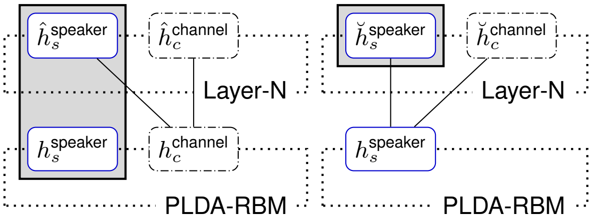
However, in cases where limited labelled data is available for back-end training, the problem of speaker information linkage to channel factors may reappear. To address this issue, we propose a deep architecture in which the channel factors of the initial PLDA-RBM are further processed using a second PLDA-RBM model. The same approach is repeated N times, leading to a deep architecture that is trained using greedy CD1. For completeness, the same idea is also applied to the speaker layers. Both approaches are described below.

3.2.1. Stacking on Channel Units

Following the hypothesis of *biometric information to be still present in hidden channel units*, the extracted hidden channel units are further examined by deeper PLDA-RBM layers, i.e. hidden channel units of the $(N\text{-}1)^{\text{th}}$-layer are re-processed by the N^{th}-layer, resulting in the hidden speaker units $\hat{h}_s^{\text{speaker}}$. Thereby, CD1 training is performed layer-wise (greedy), where L2 regularization is only applied on the first layer, since the weights of deeper layers decreased dramatically in our experimental set-up on limited training data. For stacking on channel units, we propose a feature fusion of the hidden speaker units of all layers $\{h_s^{\text{speaker}}, \dots, \hat{h}_s^{\text{speaker}}\}$ by concatenation in order to assemble an augmented reconstructed biometric feature, cf. fig 2a.

3.2.2. *Stacking on Speaker Units*

Following the hypothesis of *noisy speaker units*, the reconstructed hidden speaker units are refined by deeper PLDA-RBM layers, i.e. hidden speaker units of the $(N\text{-}1)^{\text{th}}$-layer are re-processed by the N^{th}-layer, resulting in the hidden speaker units $\breve{h}_s^{\text{speaker}}$, which we propose as biometric features, cf. fig. 2b. Though, this approach may also lead to further loss of biometric information, if the original h_s^{speaker} units comprise already well-reconstructed features, which can be over-fitted by re-assessment e.g., due to limited training data.


$\hat{h}_s^{\text{speaker}}$
$\hat{h}_c^{\text{channel}}$
Layer-N
h_s^{speaker}
h_c^{channel}
PLDA-RBM
$\breve{h}_s^{\text{speaker}}$
$\breve{h}_c^{\text{channel}}$
Layer-N
h_s^{speaker}
PLDA-RBM

(a) Stacking channel units (b) Stacking speaker units

Fig. 2. Comparison of proposed deep PLDA-RBM designs with $\hat{h}$ as deep hidden units of deeper layers N, where gray layers indicate the proposed biometric features.

4. EXPERIMENTAL ANALYSES ON MOBIO TASK

Experiments are carried out on the MOBIO SRE'13 task [7]. A standard speaker recognition front-end is used based on *rastamat* [16] and *jfacookbook* [17]: 60-dimensional speech signal features based on 19 Mel-Frequency Cepstral Coefficitans (MFCCs) with log-Energy and derived Δ and $\Delta\Delta$ coefficients on a standard hamming window. Feature warping [18] is applied using a $3\,s$ sliding window, and non-speech features are removed by unsupervised GMM voice activity detection described in [19]. Raw i-vectors are extracted with 400 dimensions based on a 512-component UBM.

4.1. Database Description: MOBIO SRE'13

The speaker recognition subset of the MOBIO database [6,7] was recorded on mobile phones and laptops, however in the MOBIO SRE'13 [7] only data from mobile phones was used. Table 1 depicts the amount of speakers and samples for each of the backround, development (dev-set) and evaluation set (eval-set). The subsets contain in total 50, 42 and 58 subjects, which is not sufficient for large-scale deep learning.

Table 1. Partitioning of MOBIO database, see [7].

Set	Female		Male	
	Subjects	Samples	Subjects	Samples
Background	13	2496	37	7104
dev-set (ref)	18	90	24	120
dev-set (prb)	18	1890	24	2520
eval-set (ref)	20	100	38	190
eval-set (prb)	20	2100	38	3990

4.2. Evaluation Criteria

The biometric performance is reported in accordance to the ISO/IEC IS 19795-1 [20] by the Equal-Error-Rate (EER), and the False Non-Match Rate (FNMR) at a 1% False Match Rate (FMR100). As an application-independent performance metric, we emphasize on the minimum cost of log-likelihood ratio (LLR) scores $C_{\text{llr}}^{\text{min}}$, which represents the generalized empirical cross-entropy of genuine and impostor LLRs with respect to Bayesian thresholds $\eta \in (-\infty, \infty)$ assuming well-calibrated systems [21,22]. As primary metric in [7], the half total error rate (HTER) is the averaged FNMR and FMR at the dev-set EER-threshold.

4.3. Baseline Systems

Systems are gender-independent due to the limited data. For the baseline G-PLDA system [2], i-vectors are projected into a spherical unit space by length-normalization, incorporating mean-subtraction, rotation by within class covariance normalization (WCCN), and projection onto the unit sphere. LDA was not applied, since no significant gains were yielded in reported systems [7].

The baseline PLDA-RBM system is based on the *Matlab Environment for Deep Architecture Learning (MEDAL)* [23] and the architecture described in Section 3. PLDA-RBM layers are CD1-trained using the background set, where the mini-batches comprise a quarter of the i-vectors per subject. Then, PLDA-RBM is re-trained using the dev-set in order to cope dataset shifts on limited short-utterance mobile data.

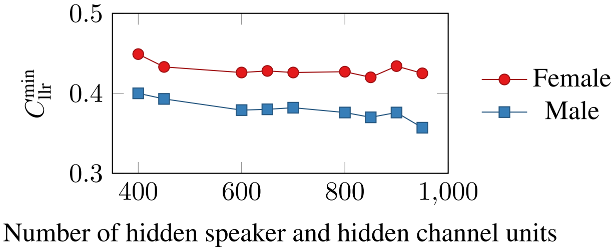
Table 2. Performance of baseline systems on dev-set.

System	Female			Male		
	EER	FMR100	$C_{\text{llr}}^{\text{min}}$	EER	FMR100	$C_{\text{llr}}^{\text{min}}$
G-PLDA [2]	15.3	63.5	0.488	**12.2**	**44.6**	**0.413**
PLDA-RBM						
GG [5]	17.7	64.2	0.552	16.7	60.4	0.526
GB	**13.5**	**51.2**	**0.451**	12.3	48.3	0.418

Table 2 indicates the baseline performance in terms of EER (in %), FMR100 (in %) and $C_{\text{llr}}^{\text{min}}$ of G-PLDA with 400 speaker factors and PLDA-RBM with 400 hidden speaker units with GG and GB energy functions. Re-training is not applied at this stage. GB PLDA-RBM significantly outperforms GG PLDA-RBM, where also performance gains to the G-PLDA baseline can be observed. However, this observation may change on big data background sets, such as on NIST SRE scenarios.

An optimal configuration regarding the number of hidden speaker and channel units was examined on dev-set. Fig. 3 depicts $C_{\text{llr}}^{\text{min}}$ for GB PLDA-RBM with dev-set re-training, where good results were observed at 850 hidden units.

Fig. 3. Comparison of different number of hidden speaker and channel factors incorporating dev-set re-training.



4.4. Experimental Results

The hypotheses of Section 3 are examined on up to three layers on dev-set, cf. table 3. While stacking on hidden speaker units decreases information in terms of $C_{\text{llr}}^{\text{min}}$, stacking on channel units is able to retrieve information.

Table 3. $C_{\text{llr}}^{\text{min}}$ comparison of stacking concepts for hidden speaker unit extraction on up to three layers on dev-set: channel units (chn-stack) and speaker units (spk-stack).

# layers	Female		Male	
	chn-stack	spk-stack	chn-stack	spk-stack
1	0.420		0.370	
2	**0.392**	0.481	**0.341**	0.452
3	0.394	0.487	0.346	0.475

Further, we investigated on the $C_{\text{llr}}^{\text{min}}$ entropy of channel unit stacked PLDA-RBM: by comparing the performance of hidden speaker units per layer to the proposed concatenation of hidden speaker units assembled from all layers, cf. table 4. Subject information can be still retrieved on the 5th layer, but without further significant gains, on which hidden speaker units are more prone to be zero in this set-up.

Table 5 compares HTER performances of the examined GB PLDA-RBM with 850 hidden speaker and channel units to female and male systems of MOBIO SRE'13, which have been reported as competitive. Contrary to state-of-the-art systems, both systems follow the Gaussian Mixture Model vs. Universal Background Model (GMM – UBM) approach.

Table 4. $C_{\text{llr}}^{\text{min}}$ comparison of recovered i-vectors by the channel-stacked PLDA-RBM architecture on N^{th}-layer (layer) and layer-concatenated (concat.) features.

	# layers	1	2	3	4	5
Female	layer	0.420	0.505	0.551	0.691	0.715
	concat.		**0.392**	0.394	0.398	0.394
Male	layer	0.370	0.459	0.510	0.639	0.681
	concat.		0.341	0.346	**0.340**	0.342

Table 5. HTER (in %) and $C_{\text{llr}}^{\text{min}}$ comparison of best single systems of MOBIO SRE'13 on eval-set to proposed systems. None of the systems incorporates calibration.

System	Female		Male	
	HTER	$C_{\text{llr}}^{\text{min}}$	HTER	$C_{\text{llr}}^{\text{min}}$
MOBIO-female [7]	11.6	n/a	9.1	n/a
MOBIO-male [7]	12.8	n/a	**8.9**	n/a
G-PLDA [2]	16.4	0.522	9.9	0.326
GB PLDA-RBM	12.0	0.397	10.6	0.361
2-layer PLDA-RBM (channel-stacked)	**11.3**	**0.368**	9.0	**0.319**

5. CONCLUSION

In this paper, we demonstrate the applicability of PLDA-RBM for limited data mobile environment speaker recognition. PLDA-RBM benefits from GB assumptions on limited mobile data outperforming the conventional G-PLDA by reconstructing speaker features and removing channel impacts. Moreover, deep PLDA-RBM is shown to recover relevant biometric information from discarded channel units by using the proposed stacking on channel units concept.

However, compared to competitive systems of MOBIO SRE'13, which rely on GMM – UBM, the proposed system achieves no significant different results, which is rather acceptable for e.g., forensics, where processing efforts are of minor concerns, but reliable evidence is rather important. Especially on female comparisons, relying on limited training data, performance gains are observed. Future research will focus on the use of drop-outs in order to increase the robustness of the RBM training, and on examining effects on large-scale NIST SRE databases with respect to optimal configurations regarding energy functions, adequate fine-tuning mechanisms and the amount of hidden speaker and channel units per layer.

Acknowledgment

This work has been partially funded by the Center for Advanced Security Research Darmstadt (CASED) and the Hesse government (project no. 467/15-09, BioMobile).

6. REFERENCES

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

[2] D. Garcia-Romero and C. Y. Espy-Wilson, “Analysis of i-vector Length Normalization in Speaker Recognition Systems,” in *ISCA Interspeech*, 2011.

[3] S. Novoselov, T. Pekhovsky, O. Kudashev, V. Mendelev, and A. Prudnikov, “Non-linear PLDA for i-Vector Speaker Verification,” in *ISCA Interspeech*, 2015.

[4] John Hopkins University Center of Excellence, “2013 Speaker Recognition Workshop,” `http://www.clsp.jhu.edu/workshops/13-workshop/speaker-and-language-recognition/`, 2013, [Online; accessed 2015-09-23].

[5] T. Stafylakis, P. Kenny, M. Senoussaoui, and P. Dumouchel, “PLDA using Gaussian Restricted Boltzmann Machines with Application to Speaker Verification,” in *ISCA Interspeech*, 2012.

[6] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes, “Bi-Modal Person Recognition on a Mobile Phone: Using Mobile Phone Data,” *IEEE ICME Workshop on Hot Topics in Mobile Multimedia*, 2012.

[7] E. Khoury, B. Vesnicer, J. Franco-Pedroso, R. Violato, and Z. Boulkenafet et al., “The 2013 Speaker Recognition Evaluation in Mobile Environment,” in *The 6th IAPR International Conference on Biometrics*, 2013.

[8] S. Cumani and P. Laface, “Generative Pairwise Models for Speaker Recognition,” in *ISCA Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014.

[9] P.-M. Bousquet, J.-F. Bonastre, and D. Matrouf, “Exploring some limits of Gaussian PLDA modeling for i-vector distributions,” in *ISCA Odyssey 2014: The Speaker and Language Recognition Workshop*, 2014.

[10] S. J. D. Prince, *Computer Vision: Models, Learning, and Inferences*, Cambridge University Press, 2012, ISBN: 978-1-107-01179-3.

[11] P. Kenny, “Notes on Boltzmann Machines,” Tech. Rep., Centre de recherche informatique de Montréal (CRIM), 2011.

[12] D. H. Ackley and G. E. Hinton, “A Learning Algorithm for Boltzmann Machines,” *Cognitive Science*, 1985.

[13] G. E. Hinton, *Neural Networks: Tricks of the Trade*, chapter A Practical Guide to Training Restricted Boltzmann Machines, Springer, 2012, ISBN: 978-3-642-35289-8.

[14] Y. Bengio, “Learning Deep Architectures for AI,” *Foundations and trends in Machine Learning*, 2009.

[15] T. Yamashita, M. Tanaka, E. Yoshida, Y. Yamauchi, and H. Fujiyoshi, “To be Bernoulli or to be Gaussian, for a Restricted Boltzmann Machine,” in *IAPR IEEE International Conference on Pattern Recognition (ICPR)*, 2014.

[16] D. P. W. Ellis, “PLP and RASTA (and MFCC, and inversion) in Matlab,” `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`, 2005, [Online; accessed 2014-03-26].

[17] O. Glembek, “Joint Factor Analysis Matlab Demo,” `http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo`, 2009, [Online; accessed 2013-10-10].

[18] J. Pelecanos and S. Sridharan, “Feature Warping for Robust Speaker Verification,” in *ISCA Speaker Odyssey 2001: The Speaker Recognition Workshop*, 2001.

[19] J. Alam, P. Kenny, P. Ouellet an T. Stafylakis, and P. Dumouche, “Supervised/Unsupervised Voice Activity Detectors for Text-dependent Speaker Recognition on the RSR2015 Corpus,” in *ISCA Speaker Odyssey 2014: The Speaker Recognition Workshop*, 2014.

[20] ISO/IEC, “Information technology – Biometric performance testing and reporting – Part 1: Principles and framework,” ISO/IEC 19795-1:2006, JTC 1/SC 37, Geneva, Switzerland, 2011.

[21] D. Ramos and J. Gonzalez-Rodriguez, “Cross-entropy Analysis of the Information in Forensic Speaker Recognition,” in *ISCA Odyssey 2008: The Speaker and Language Recognition Workshop*, 2008.

[22] N. Brümmer and J. du Preez, “Application-Independent Evaluation of Speaker Detection,” in *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006.

[23] D. E. Stansbury, “Matlab Environment for Deep Architecture Learning (MEDAL) v0.1,” `https://github.com/dustinstansbury/medal`, 2013, [Online; accessed 2015-09-22].