DIVERGENCE ESTIMATION BASED ON DEEP NEURAL NETWORKS AND ITS USE FOR LANGUAGE IDENTIFICATION

Yosuke Kashiwagi, Congying Zhang, Daisuke Saito, Nobuaki Minematsu

The University of Tokyo, Japan

{kashiwagi, zhangcy, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

In this paper, we propose a method to estimate statistical divergence between probability distributions by a DNN-based discriminative approach and its use for language identification tasks. Since statistical divergence is generally defined as a functional of two probability density functions, these density functions are usually represented in a parametric form. Then, if a mismatch exists between the assumed distribution and its *true* one, the obtained divergence becomes erroneous. In our proposed method, by using Bayes' theorem, the statistical divergence is estimated by using DNN as discriminative estimation model. In our method, the divergence between two distributions is able to be estimated without assuming a specific form for these distributions. When the amount of data available for estimation is small, however, it becomes intractable to calculate the integral of the divergence function over all the feature space and to train neural networks. To mitigate this problem, two solutions are introduced; a model adaptation method for DNN and a sampling approach for integration. We apply this approach to language identification tasks, where the obtained divergences are used to extract a speech structure. Experimental results show that our approach can improve the performance of language identification by 10.85% relative compared to the conventional approach based on i-vector.

Index Terms— language identification, deep neural network, i-vector, statistical divergence, structural feature

1. INTRODUCTION

Statistical divergence between distributions, such as Kullback-Leibler (KL) divergence or Bhattacharyya divergence (BD), has been widely used in machine learning. In general, since statistical divergence is defined as a functional of two probability density functions, a parametric form of the distribution is required to calculate the divergence. Since the true distribution can have a complex shape and is often difficult to estimate precisely, a simple distribution like the Gaussian distribution is insufficient. One of the approaches to increase estimation accuracy is to apply a more complex model to the feature distribution. For example, Gaussian Mixture Models (GMMs) are adopted for modeling and its effectiveness is shown in [1, 2]. However, if a mismatch still exists between the true distribution and GMMs, it also causes estimation errors of the divergence.

Recently, Deep Neural Networks (DNNs) have become one of the main streams of acoustic modeling, and a variety of leaning techniques have been proposed [3, 4]. Since DNNs are powerful models, some generative approaches are replaced with discriminative approaches such as DNN-HMM in acoustic modeling. In discriminative models, the parametric form of the feature distribution is not explicitly assumed. Hence, these models are more flexible than generative ones. In this paper, DNN-based flexibility is introduced to the process of estimating the statistical divergence.

Heigold *et al.* proposed a method of estimating statistical divergence using the log-linear model, which is a typical model of discriminative ones [5]. In their approach, the parameters of the log-linear model are converted into those of the Gaussian distribution, then estimation of the divergence was done as calculation of these converted parameters. Applying this approach, Li *et al.* proposed a training algorithm for DNNs, which uses the estimated divergence as criteria from the softmax layer of the DNNs [6]. However, since the converted parameters of Gaussian distribution in the above approach include uncertainties, which are explained shortly, some restrictions such as shared covariance matrices are required. In addition, although this approach is based on a discriminative manner, use of Gaussian distributions indicates inevitable assumption of the shape of distribution. Hence, the accuracy of the estimated divergence is influenced by the degree of mismatch between the adopted generative models and the *true* distribution.

This paper describes a new discriminative technique to estimate the statistical divergence not using generative model parameters at all. We introduce Bayes' theorem to the BD, rewrite its well-known analytical form into another form using posterior terms, and estimate the posterior terms by DNNs. To evaluate the proposed method, we carry out language identification experiments using the estimated divergence as structural features.

This paper is organized as follows. Section 2 introduces the conventional method that uses the DNNs to estimate divergence with the log-linear model. We formulate the proposed approach in Section 3 and describe our language identification system in Section 4. Experimental results are given in Section 5. Finally our paper is concluded in Section 6.

2. CONVENTIONAL APPROACH USING LOG-LINEAR MODEL

Heigold *et al.* proposed a method which estimates the statistical divergence using the log-linear model. First, the parameters of the model are converted into the parameters of Gaussian distribution. Next, the statistical divergence such as KL divergence is calculated by the closed form using the converted parameters.

The log-linear model can estimate the posterior of label y given input feature vector $\boldsymbol{x}$ as:

$$p(y|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left(\sum_i \lambda_{yi} f_i(\boldsymbol{x})\right), \tag{1}$$

where i represents the order, f_i is the generalized feature, weights of $\{\lambda_{yi}\}$ are the set of model parameters and $Z(\boldsymbol{x})$ is the normalization

factor. Softmax function, which is used in DNNs as the output layer, can be interpreted as the log-linear model as:

$$\begin{aligned} f_1(\boldsymbol{x}) &= [\boldsymbol{x}^\top, 1]^\top, \\ \lambda_{y1} &= [\boldsymbol{W}_y, b_y], \\ \lambda_{yi} &= 0 \quad (i \geq 2). \end{aligned} \tag{2}$$

When the distribution of features corresponding to each label is assumed to be Gaussian distribution, its parameters can be derived from the parameters of softmax function as:

$$\Sigma_y = \Sigma, \qquad \boldsymbol{\mu} = \Sigma^{-1}[\boldsymbol{W}_y, b_y], \tag{3}$$

where Σ is the shared covariance matrix for all labels.

Li *et al.* use this approach to train small scale DNNs from large scale ones. They used the statistical divergence as training criteria, and achieved a performance improvement. However, this approach still approximates the feature distribution as Gaussian distribution. In addition, according to equation 3, the mean and covariance parameters include the uncertainty of the scale.

3. DISCRIMINATIVE APPROACH FOR ESTIMATION OF THE STATISTICAL DIVERGENCE USING DNNS

As told in the previous section, if there are mismatches between the assumed distribution and its true distribution, the obtained distribution becomes erroneous. Hence, we can say that it is more adequate to estimate the divergence not using a generative model. To address this problem, we estimate the statistical divergence using DNNs as discriminative model through Bayes' theorem. In this study, we adopt the BD as the statistical divergence. It is defined as:

$$BD(a,b) = -\ln \int \sqrt{p(\boldsymbol{x}|y=a)p(\boldsymbol{x}|y=b)}d\boldsymbol{x} \ , \tag{4}$$

where a and b denote classes of acoustic models, such as phoneme states. When the distribution of features corresponding to each label is assumed to be Gaussian distribution, the BD can be calculated as:

$$\begin{aligned} BD(a,b) &= \frac{1}{8}(\boldsymbol{\mu}^{(a)} - \boldsymbol{\mu}^{(b)})^\top \Sigma^{-1}(\boldsymbol{\mu}^{(a)} - \boldsymbol{\mu}^{(b)}) \\ &\quad + \frac{1}{2}\ln\left(\frac{\det \Sigma}{\sqrt{\det \Sigma^{(a)} \det \Sigma^{(b)}}}\right), \\ \Sigma &= \frac{\Sigma^{(a)} + \Sigma^{(b)}}{2}, \end{aligned} \tag{5}$$

where $\boldsymbol{\mu}^{()}$ and $\Sigma^{()}$ denote the parameters of Gaussian distributions.

In the case that DNN-based acoustic models are available, they can directly calculate posterior probabilities such as $p(y=a|\boldsymbol{x})$ and $p(y=b|\boldsymbol{x})$ Applying Bayes' theorem to Eqaution (4), the BD is represented as a functional of the posterior probabilities as:

$$\begin{aligned} BD(a,b) &= -\ln \int \sqrt{p(\boldsymbol{x}|y=a)p(\boldsymbol{x}|y=b)}d\boldsymbol{x} \\ &= -\ln \int \sqrt{\frac{p(y=a|\boldsymbol{x})p(\boldsymbol{x})}{p(y=a)}\frac{p(y=b|\boldsymbol{x})p(\boldsymbol{x})}{p(y=b)}}d\boldsymbol{x} \\ &= -\ln \int p(\boldsymbol{x})\sqrt{p(y=a|\boldsymbol{x})p(y=b|\boldsymbol{x})}d\boldsymbol{x} \\ &\quad + \frac{1}{2}\ln p(y=a) \\ &\quad + \frac{1}{2}\ln p(y=b). \end{aligned} \tag{6}$$

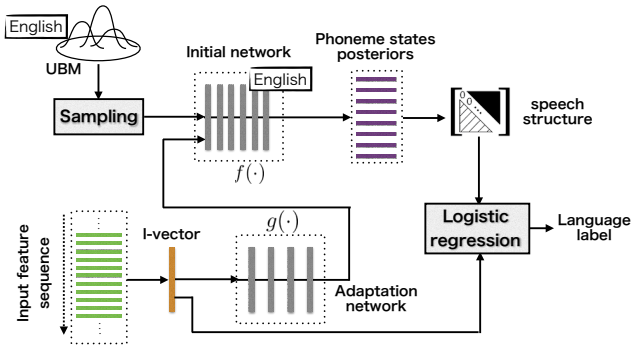

Fig. 1. The proposed language identification system.

Thus, it is possible to calculate the BD not via generative model parameters. Though we adopt BD as the statistical divergence, other statistical divergence can be similarly adopted in this approach.

4. LANGUAGE IDENTIFICATION USING STRUCTURAL FEATURES

4.1. Overview

When one wants to estimate the BD by equation (6), we can say that two issues have to be handled carefully. Posteriors can be calculated by DNNs but the DNNs are difficult to be trained only with a small amount of data like one or a few utterances. The other issue is related to integration over the entire space in equation (6). Only a small amount of data, which are available for language identification, will occupy only a portion of the space, i.e., biased data. It is easily assumed that some phonemes are missing in a few utterances and this will surely increase errors. To address these issues, we use adaptation and sampling approaches.

Fig. 1 shows the flowchart of our language identification system. In this task, since language identity is unknown, the system assumes at first that all the input utterances as English utterances. Then, using equation (6), it calculates the BD between every possible pair of the 132 English phoneme states. For calculation, DNNs, which are used as English phoneme posterior estimator, are adapted to the input utterance. Further, for the integral over the entire space in equation (6), we use a sampling method for a spoken English corpus, WSJ, which was used to train the initial DNNs. It is highly expected that the BD between phoneme states depends on input utterances and they will be used as additional features for the task of language identification. The additional features are called structural features in this paper.

The detailed procedure of language identification is as follows. First, the utterance-wise i-vector is extracted from each utterance. The extractor is trained using the data from multiple languages (NIST LRE [7]). Next, the initial DNNs are trained by the English corpus of WSJ and they are adapted to the utterance by using the extracted i-vector [8]. After that, the statistical divergences between all the phoneme states are estimated by equation (6) using the sampling approach and the adapted DNNs. In the sampling step, the universal background model (UBM) of WSJ is used for sampling so that sampled observations can cover the entire space. Finally the language label for each of input utterances is estimated from its i-vector and its structural features constructed by the BDs of all the phoneme states. In the following sections, the details of adaptation and sampling are described.

4.2. Adaptation of DNNs

In the adaptation step, we use an utterance-based adaptation method which was proposed by Miao *et al.* [8]. This approach uses adaptation network $g(\cdot)$ (See Fig. 1) which estimates the bias factor from i-vector and the bias is used as input to $f(\cdot)$. Then, $f(\cdot)$ is adapted to the input utterance and it calculates posterior probabilities.

$$\begin{aligned} p(\boldsymbol{y}_t|\boldsymbol{o}_t) &= f(\boldsymbol{a}_t), \\ \boldsymbol{a}_t &= \boldsymbol{o}_t + g(\boldsymbol{i}_s), \end{aligned} \tag{7}$$

where $\boldsymbol{o}_t$ is input feature vector, $\boldsymbol{y}_t$ denotes the label and $\boldsymbol{i}_s$ is an i-vector. In our language identification system, the i-vector is extracted using the total variability matrix calculated from multiple languages. So, the extracted i-vector is supposed to carry language attributes rather than speaker ones.

The brief overview of the training strategy of the two networks of $f(\cdot)$ and $g(\cdot)$ is as follows. 1) train global DNNs $f(\cdot)$, 2) connect the adaptation network $g(\cdot)$ to $f(\cdot)$ and update the parameters of $g(\cdot)$ while keeping the parameters of $f(\cdot)$, and 3) update the parameters of $f(\cdot)$ again while keeping the parameters of $g(\cdot)$.

As noted above, phoneme state labels are required to construct the DNN-based acoustic model. However, it is theoretically impossible to apply the set of phonemes of any single language to multiple languages. Then, in this work, the DNNs were trained using WSJ (English corpus) only. Therefore, non-English utterances are treated as utterances of English with a very unique accent. The structural features, a full set of the BDs between phoneme states, are supposed to characterize the uniqueness of each accent, i.e., language.

4.3. Estimation of phoneme posterior vectors using a sampling approach

With the adapted DNNs, the statistical divergence of equation (6) is estimated by a sampling approach as:

$$\begin{aligned} BD(a,b) &= -\ln \frac{1}{N} \sum_n \sqrt{p(y_n = a|\boldsymbol{x}_n, \theta_c) p(y_n = b|\boldsymbol{x}_n, \theta_c)} \\ &\quad + \frac{1}{2} \ln \frac{1}{N} \sum_n p(y_n = a) \\ &\quad + \frac{1}{2} \ln \frac{1}{N} \sum_n p(y_n = b), \end{aligned} \tag{8}$$

where N is the number of features in an observed utterance, $\{\boldsymbol{x}_n\}$ denotes the set of features, and θ_c is the set of the parameters of the networks which were adapted to that observed utterance.

If we use a set of feature samples collected only from the observed utterance, however, it is intractable to calculate the summation over all the feature space in equation (8). To address this problem, we effectively use UBMs for integration and equation (8) will be slightly modified into equation (9):

$$\begin{aligned} BD(a,b) &= -\ln \frac{1}{M} \sum_m \sqrt{p(y_m = a|\boldsymbol{x}_m, \theta_c) p(y_m = b|\boldsymbol{x}_m, \theta_c)} \\ &\quad + \frac{1}{2} \ln \frac{1}{L} \sum_l p(y_l = a) \\ &\quad + \frac{1}{2} \ln \frac{1}{L} \sum_l p(y_l = b), \end{aligned} \tag{9}$$

where M is the number of samples generated from the UBMs and $\{\boldsymbol{x}_m\}$ denotes the set of samples. L is the number of the training data for the UBMs, i.e., WSJ. The prior terms in equation (8) can be approximated as constant values as:

$$\begin{aligned} \frac{1}{N} \sum_n p(y_n = a) &\approx \frac{1}{L} \sum_l p(y_l = a), \\ \frac{1}{N} \sum_n p(y_n = b) &\approx \frac{1}{L} \sum_l p(y_l = b). \end{aligned} \tag{10}$$

The two left terms correspond to alignment of each phoneme state to observed features.

4.4. Structured features and identification models

After estimating the statistical divergence using DNNs, we construct structural features, which are a full set of the BDs between every possible phoneme state pair [9]. The dimension of the features is $K(K-1)/2$ where the number of the phoneme states is K. In studies of automatic speech recognition, changes of cepstrum coefficients caused by differences of sex and age are well characterized as affine transformation [10].

If we use Gaussian distribution to model the feature distribution of each phoneme state, the BD is invariant to any static affine transformation. This means that the structural features are robust to speaker differences [11].

In our system, we estimate the distributions without assuming any specific form. If the estimated distribution can fit the true one very well, the divergence achieves perfect robustness to all bijective transformations[1] [12]. That is to say, the estimated divergence could be insensitive to 'language variation' due to perfect robustness. In this case, it is useless for language identification. However, since the DNNs for divergence estimation are actually trained only by a specific language, robustness is expected to be weakened. For that, with language variation, the divergence show different values. Hence it is expected that use of the structural features enhances the performance of language identification as the features were effectively introduced to accent clustering [13].

As was shown in section 4.1, even if all the kinds of phoneme states are not observed in input utterances, our system can calculate the divergence only from i-vector. Finally, the structural features and i-vector are used for logistic regression as input.

5. EXPERIMENTS

The performance of our language identification system was evaluated using NIST LRE 2007 database. The language-closed condition was adopted and the number of languages is 14. Here, the evaluation set had data of 3 types of duration (3, 10 and 30 seconds). For training the i-vector extraction model and the logistic regression model, we used NIST LRE 2003, NIST LRE 2005, NLST LRE 2007. On the other hand, only WSJ was used for training the initial DNNs. As explained before, the UBM for sampling was also trained with WSJ.

For calculation of i-vectors, 6-dimensional MFCCs and power were used and the dimension of the i-vector was 600. As input to the DNNs, a central frame of 12-dimensional MFCCs and C0, and its neighboring 10 frames were used. The number of hidden nodes in each layer of the DNNs for posterior estimation was 1024, and the number of hidden layers was 6. For the DNNs for adaptation, the number of hidden nodes in each layer was also 1024, but the number of hidden layers was set to 4. The output phoneme states

[1] The BD originally and theoretically has perfect robustness to all bijective transformations.

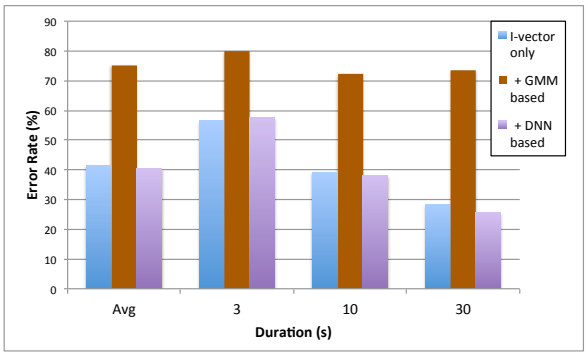

Fig. 2. Comparison among our proposed system and the two baseline systems in terms of error rates

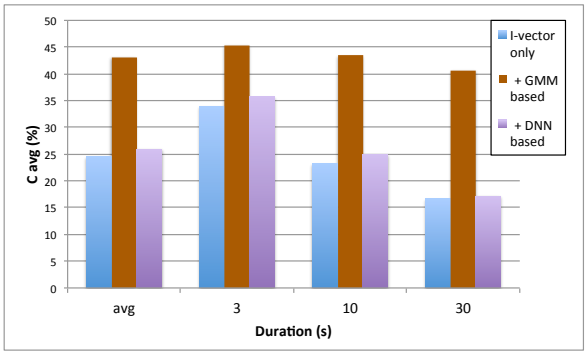

Fig. 3. Comparison among our proposed system and the two baseline systems in terms of average costs

were labeled using GMM-HMM monophone models which had 132 phoneme states, so that the dimension of structural features was 8,646. The number of mixtures of UBMs was 1024 and the number of frames used for sampling varied from 10 to 1000 for testing.

First we compare our proposed system which uses 100 samples with two baseline systems. One uses the i-vector estimated from an input utterance and only the i-vector is used for language identification. The other estimates a GMM from an input utterance and calculates the BD between every possible Gaussian pair. The resulting structural features as well as the i-vector are jointly used for language identification.

In the latter baseline system, the GMM was adapted from the UBM-GMM using MAP-adaptation and it had 128 mixtures, so the structural features here had 8,128 dimensions. Fig. 2 and Fig. 3 show the language identification error rates and the average costs (C_{avg}) as a function of the duration of input utterances for each of the two baseline systems and our proposed system.

Fig. 4 and Fig. 5 show the language identification error rates and average costs (C_{avg}) as a function of the duration of input utterances for different numbers of sampled frames. Baseline here was the performance of the i-vector system. If the number of sampled frames is small such as 10 and 50, the performance of our approach is lower than that of the baseline system. However, it is clearly shown that our proposed approach becomes effective when the number increases up to 1,000. This is considered to be because the estimation accuracy was not good enough when the number of sampled frames was small.

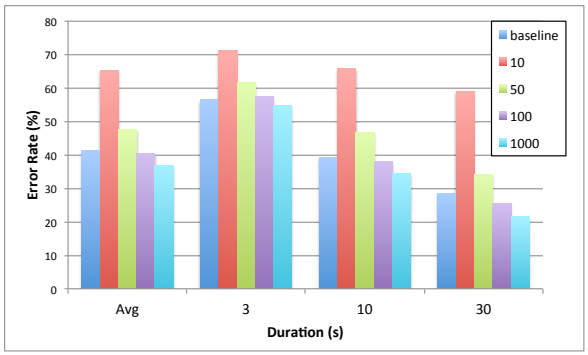

Fig. 4. Error rates as a function of the duration of input utterances for different numbers of sampled frames

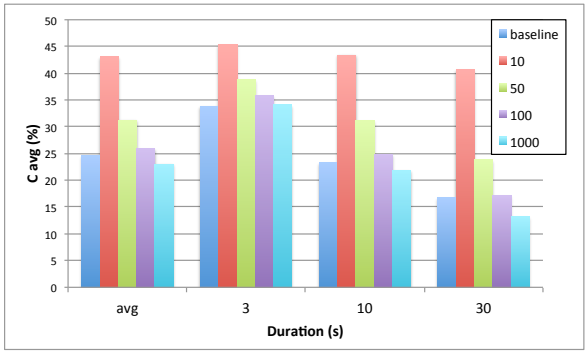

Fig. 5. Average costs as a function of the duration of input utterances for different numbers of sampled frames

6. CONCLUSION

In this paper, we proposed a method to estimate the statistical divergence between feature distributions without assuming any specific form of the distributions. This estimation was realized by introducing DNNs as discriminative model. The resulting divergences are used for the task of language identification. Here, the divergences are utilized as structural features. Two issues were carefully handled: only a very limited amount of utterances are available to build the DNNs and the integral operation over the entire space has to be run to estimate the divergence. We exploited adaptation and sampling for the two issues, respectively. Experiments showed that our system can reduce the identification error rate by 10.85% relative.

7. REFERENCES

[1] John R Hershey, Peder Olsen, et al., "Approximating the Kullback Leibler divergence between gaussian mixture models," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, pp. 317–320.

[2] Chang Huai You, Kong Aik Lee, and Haizhou Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1300–1312, 2010.

[3] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[4] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[5] Georg Heigold, Hermann Ney, Patrick Lehnen, Tobias Gass, and Ralf Schlüter, "Equivalence of generative and log-linear models," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1138–1148, 2011.

[6] Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong, "Learning small-size DNN with output-distribution-based criteria," in *Proc. Interspeech*, 2014, pp. 1910–1914.

[7] NIST LRE Group, "The 2007 nist languag recognition evaluation plan (lre07)," 2007.

[8] Yajie Miao, Lu Jiang, Hao Zhang, and Florian Metze, "Improvements to speaker adaptive training of deep neural networks," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, 2014, pp. 165–170.

[9] Nobuaki Minematsu, "Yet another acoustic representation of speech sounds," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, 2004, vol. 1, pp. 585–588.

[10] Michael Pitz, Sirko Molau, Ralf Schlüter, and Hermann Ney, "Vocal tract normalization equals linear transformation in cepstral space.," in *INTERSPEECH*, 2001, pp. 2653–2656.

[11] Nobuaki Minematsu, Satoshi Asakawa, Masayuki Suzuki, and Yu Qiao, "Speech structure and its application to robust speech processing," *New Generation Computing*, vol. 28, no. 3, pp. 299–319, 2010.

[12] Yu Qiao and Nobuaki Minematsu, "A study on invariance of f-divergence and its application to speech recognition," *Signal Processing, IEEE Transactions on*, vol. 58, no. 7, pp. 3884–3890, 2010.

[13] Nobuaki Minematsu, Shun Kasahara, Makino Takehiko, Saito Daisuke, and Hirose Keikichi, "Speaker-basis accent clustering using invariant structure analysis and the speech accent archive," *Proc. Odyssay*, pp. 158–165, 2014.