Investigation of Cross-lingual Bottleneck Features in Hybrid ASR Systems

Jie Li, Rong Zheng, Bo Xu

Interactive Digital Media Technology Research Center,
Institute of Automation, Chinese Academy of Sciences, Beijing, P.R.China
{jie.li, rong.zheng, xubo}@ia.ac.cn

Abstract

In this work, we investigate cross-lingual BN features in hybrid ASR systems under a two-level DNNs framework. The first-level DNNs are bottleneck feature extractors and the second-level DNNs serve as not only acoustic models but also feature combination modules. Different feature configurations, including the bottleneck dimensionality, the need of delta processing and the necessity of concatenation with standard features of target-language, are first studied. Further experiments are done to evaluate the cross-lingual generalization in a more holistic manner using optimized features. We then analyze the effects of adding more training data on the BN feature extractors. Performance improvement can be obtained when more data available. Finally, two different approaches of utilizing data from non-target languages are experimentally compared. It is shown that these two approaches have similar performance with each other, and the two-level DNNs architecture benefits from either of them.

Index Terms: cross-lingual, bottleneck features, hybrid systems, deep neural networks

1. Introduction

Nowadays, there are two main approaches to incorporate neural networks (NNs) in acoustic modeling: the hybrid system and the tandem system. For the former, neural networks are trained to directly estimate posterior probabilities which can then be used as scaled likelihood estimates for the states of Hidden Markov Models (HMMs). In tandem frameworks, however, neural networks act as feature extractors which are considered as non-linear feature transformation. The traditional Gaussian Mixture Models (GMMs) and HMMs are trained on standard acoustic features concatenated with discriminative features derived from neural networks, which are either the estimated class posterior probabilities (probabilistic features [1]) or the linear outputs of the neurons in the bottleneck layer (bottleneck features [2]).

In state-of-the-art automatic speech recognition (ASR) systems, acoustic models are typically trained using large amount of language-specific transcribed speech data. However, for languages with limited training data, the performance is rather low since only small models with poor accuracy can be estimated. To alleviate this problem, several techniques has been investigated which are able to generalize across languages and efficiently use data from others to boost performance on the target one. The transcribed data from other languages can be utilized to build multilingual acoustic models [3, 4], such as multilingual Subspace Gaussian Mixture Models (SGMMs) [5, 6] and multilingual Deep Neural Networks (DNNs) [7, 8]. These methods perform data borrowing on the model level. As an alternative approach, Multilayer Perceptron (MLP) can be trained and serves as data-driven feature front-ends. Cross-lingual and multilingual research indicates that MLP features (probabilistic features and BN features) possess the language independent property to a certain degree [9, 10, 11]. Many experimental results show that features extracted from an MLP which is trained with one language can be used for another one. For example, [10] studied the performance of probabilistic and BN features on different language than they were trained for. It was shown that this cross-lingual porting is possible and the features are still competitive to PLP features. In [11], cross-lingual portability of long-term bottleneck features was investigated in concatenation with MFCC, and it showed that the topology of the NN is more important than the training language, since almost all NN features achieve similar performance. The results in [12] showed that multilingual BN features consistently outperform monolingual BN features and the NN can produce very good BN features even for unseen languages.

The works listed above are all situated in the tandem framework. In this paper, however, we investigate cross-lingual BN features in hybrid ASR systems. BN feature front-ends are combined with the following Deep Neural Networks (DNNs) acoustic model. In this two-level DNNs framework, the first-level bottleneck DNNs (BN-DNNs) are trained using data from non-target languages and act as cross-lingual BN feature extractors. The BN features are then augmented with standard features of target language and sent into the acoustic model DNNs (AM-DNNs) in the second level, which serves also as feature combination modules. Under this framework, we first experimentally study the impacts of different feature configurations, including the bottleneck dimensionality, the need of delta processing and the necessity of concatenation with standard features of target-language. Then, the cross-lingual generalization of BN features in hybrid ASR systems is evaluated in a more holistic manner. Furthermore, the influence of adding more training data on BN-DNNs is analyzed. At the end, we compare different approaches of utilizing data from multi-languages.

The remainder of the paper is organized as follows: After an overview of the related work in Section 2, Section 3 describes the two-level DNNs framework in detail. A short description of the training and testing corpora is given in Section 4 and the experimental setups are introduced in Section 5. We report our experimental results in Section 6 and conclude this work in Section 7.

2. Relation to prior work

While cross-lingual BN features have been fully studied under the tandem framework in [9, 10, 11, 12], only few works focusing on applying BN features in hybrid systems have been published.

In [13], the authors investigated whether BN features are

useful for DBN/HMM hybrids for the first time. The layers of BN networks were fixed and applied to successive windows of feature frames in a time-delay fashion. Experimental results showed that BN features can improve the recognition performance of hybrid systems. The modular combination proposed in [13] was extended by [14] in multi-lingual settings. It was shown that DNN acoustic models can benefit significantly from BN features trained on different languages. However, in [13, 14], acoustic modeling was operated only on the outputs of feature front-end networks, without the original features, which may limit the power of the DNN acoustic models.

The Multi-level Adaptive Networks (MLAN) architecture, which was presented in [15] and further extended in [16, 17], can effectively combine in-domain and out-of-domain (OOD) training data. The advantage of the MLAN approach is that the second-level DNNs are able to discriminatively select the most important elements of the OOD features which are extracted by the first-level NNs.

Our approach has a similar architecture with MLAN. However, instead of studying feature combination in a cross-domain setting, we move to cross-lingual environments and focus on cross-lingual BN features in hybrid ASR systems. In [13, 14], BN networks were applied in a time-delay fashion, but in this work, concatenated frames are sent into the networks timely, which may make the system more convenient to use in practice. In addition, to make full use of DNN acoustic model, BN features are augmented with standard features.

3. Framework description

The proposed two-level DNNs framework is illustrated in Figure 1.

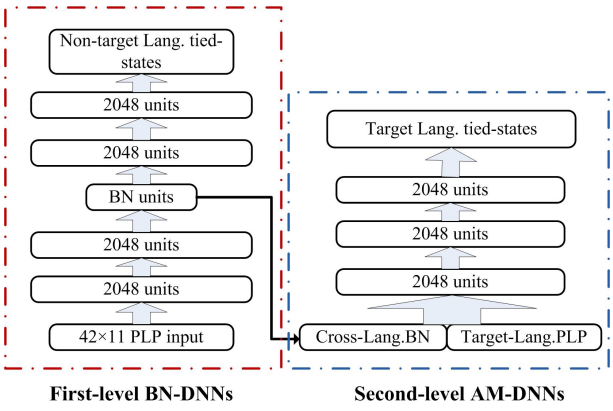

Figure 1: *Architecture of the two-level DNNs framework.*

BN-DNNs in the first level contain five hidden layers and the bottleneck layer is placed symmetrically in the model structure. To get better features, BN-DNNs are trained to predict tied states of context-dependent triphone HMMs rather than the monophone states [18]. After pre-trained and fine-tuned on non-target language data, BN-DNNs are used to process target-language acoustic data to generate cross-lingual BN features, which are then concatenated with the original target-language acoustic features and sent into the second-level AM-DNNs which serve as not only acoustic models but also feature combination modules. RBM-based pretraining for AM-DNNs is also performed on the augmented BN features.

4. Language resources

The language resources in this work come from CallHome corpora collected by Linguistic Data Consortium (LDC). For our research purpose, the following six languages are used: Spanish (SP), Mandarin (MA), English (EN), Arabic (AR), German (GE) and Japanese (JP). For each language, a phone set and phonetic lexicon are also supplied in this corpora. The detail information of the corpus is listed in Table 1.

Table 1: *Training and testing corpora information for the six languages.*

Language	Training hours	Testing hours	#phn	lexicon size	PPL
Spanish	16.5	1.9	29	46K	135
Mandarin	15.6	1.5	38	44K	167
English	14.9	1.8	43	91K	108
Arabic	13.6	1.4	41	51K	195
German	14.7	1.8	46	319K	119
Japanese	15.1	2.1	34	135K	73

For each language, a 2-gram language model (LM) is estimated with only the transcripts of specific language training data and smoothed by the modified Kneser-Ney method. The lexicon sizes and perplexity (PPL) values of unpruned LMs measured on testing sets are also listed in Table 1.

5. Experimental setup

5.1. Training of GMM-HMMs

For each language, standard tied-state cross-word triphone GMM-HMMs are first trained with maximum likelihood estimation (MLE) using the regular 42-dimension features, which consist of 13-dimensional PLP and smoothed F0 appended with the first and second order derivatives. Each of the six models contains about 1600 tied triphone states, with an average of 16 Gaussian components per state. Labels at frame level generated by GMM-HMMs are used for network fine-tuning on both BN-DNNs and AM-DNNs for each language.

5.2. Training of BN-DNNs

All BN-DNNs used in this work contain 5 hidden layers with 2048 units in each non-bottleneck layer and 20-60 units in the BN layer considering different configurations. 11-frames of 42-dimensional PLP features are used as input features for the networks. With weights initialized by pre-trained RBMs, BN-DNNs are fine-tuned to classify the tied-states of triphone for specific language. After training done, BN-DNNs are treated as feature extractors. When training and testing languages differ, the resulting BN features are referred as cross-lingual features, and intra-lingual BN features are produced when training and testing languages match.

5.3. Training of AM-DNNs

In this study, AM-DNNs are fixed to the following structure: 3 hidden layers with 2048 units per layer. The BN features extracted from BN-DNNs are used alone or concatenated with original features (11-frames of 42-dim PLP features, *11-PLP for short*) to train AM-DNNs. The pre-training of RBMs for AM-DNNs is also conducted on the corresponding features. Fine-tuning process of AM-DNNs is the same with that for BN-

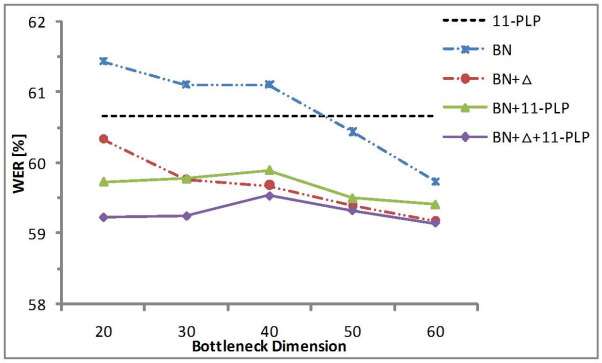

Figure 2: *Performance of different configurations for intra-lingual BN features.*

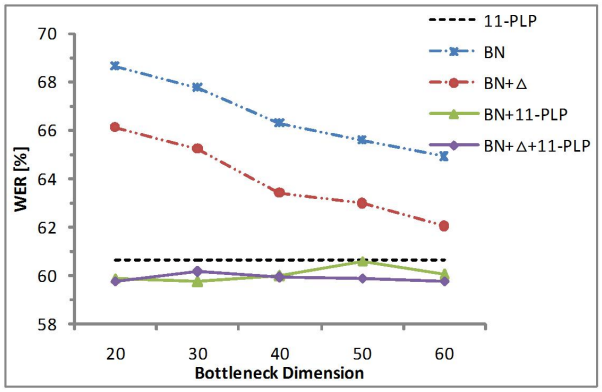

Figure 3: *Performance of different configurations for cross-lingual BN features.*

DNNs, which is performed using stochastic gradient descend (SGD) algorithm with cross-entropy criterion.

6. Experimental results

6.1. Optimizing BN features

BN features have been optimized in [19] under the tandem framework. To get the optimal parameters in hybrid ASR systems, different configurations of BN features are experimentally studied in both intra- and cross-lingual settings. We focus on three aspects: the bottleneck dimensionality, the need of delta processing and the necessity of concatenation with original features.

Spanish and English are chosen as target and non-target languages respectively. For intra-lingual experiments, BN-DNNs and AM-DNNs are both trained with Spanish training data. Figure 2 shows the performance of AM-DNNs trained using different configurations of intra-lingual BN features.

As for cross-lingual scenario, BN-DNNs are first trained with English training data and then act as feature front-ends to process Spanish data to extract cross-lingual BN features, which are used alone or augmented with Spanish original features (11-PLP) to train AM-DNNs. The results are shown in Figure 3. These two settings share the same baseline results, which are obtained from AM-DNNs trained using Spanish training data with 11-frames of 42-dim PLP features.

We make several observations from these two figures. **First of all**, for both intra- and cross-lingual settings, when *BN* or *BN+△* are used, WER is decreasing with the increasing of feature size. This means that more information is passed from BN-DNNs to AM-DNNs. However, this trend becomes unstable when combined features are used. The redundant information contained in BN and PLP features may make the system more sensitive to the bottleneck dimension. **Secondly**, delta processing is important for both intra- and cross-lingual BN features. The intra-lingual results show that *BN+△* performs even better than *BN+11-PLP* though the 42-dim PLP features also contain delta information. The bottom two lines in Figure 3 indicate that, when cross-lingual BN features are combined with PLP features, delta processing can make the system more stable, though it may not bring significant performance gains. **Thirdly**, compared with intra-lingual settings, concatenation with original features is much more necessary for cross-lingual BN features. In the intra-lingual experiments, *BN+△* performs almost the same with *BN+△+11-PLP* when the feature dimension is bigger than 50. For cross-lingual BN features, however, only when concatenated with target-language PLP features, can the system perform better than the baseline. This is mainly due to that target-language PLP features contain language specific information which can not be obtained by cross-lingual BN features.

According to the results mentioned above, the feature configuration *BN+△+11-PLP* is chosen for all the following experiments. The bottleneck dimension is fixed to 50 considering the tradeoff between efficiency and performance.

6.2. Cross-lingual generalization

The cross-lingual generalization of BN features is evaluated in this section. All of the six languages act as target and non-target language in turn. As can be seen from the results in Table 2, for each language, both the augmented intra- and cross-lingual BN features perform better than 11-PLP baseline system. This indicates that BN features extracted from the first-level BN-DNNs can be effectively combined with original acoustic features by the second-level AM-DNNs, and these two kinds of features provide complementary information.

Another noticeable thing is, that for one particular target language, no matter which one of other 5 languages is chosen to train BN-DNNs, cross-lingual BN features perform almost the same with each other and only slightly worse than the corresponding intra-lingual BN features. It can be deduced from this observation, that common information exists in speech data across different languages and each BN-DNNs have similar ability to capture this commonalities. This point ensures the generalization capability of BN features.

6.3. Effects of more training data

It should be noticed that all the six BN-DNNs in Section 6.2 are trained on similar amount of speech data. This makes us to wonder whether BN-DNNs can capture more information when more training data are included. Spanish and Mandarin are chosen as two totally different target languages and English is used as non-target language. To simulate the situation, several training sets with different data volume are randomly selected from Switchboard (SWB) corpus, which has the similar telephone conditions with CallHome. Each set is then combined with CallHome English (CH-EN) data to train BN-DNNs. To isolate the effectiveness of adding more training data, three things are determined. **First of all**, small SWB sets are included in bigger SWB sets. For example, SWB-50h contains the data of

Table 2: *Coss-lingual generalization of BN features; comparison with 11-PLP baseline.*

BN-DNNs Train Language	Test Target Language [WER %] Spanish	Mandarin	English	Arabic	German	Japanese
Spanish	**59.3**	62.6	49.6	60.9	58.3	59.0
Mandarin	59.8	**62.0**	49.9	61.2	58.2	59.3
English	59.9	62.5	**49.4**	61.2	58.3	59.0
Arabic	60.0	62.8	50.1	**60.4**	58.6	59.3
German	60.3	62.8	49.8	61.2	**57.6**	59.3
Japanese	60.2	62.4	49.9	61.3	58.4	**58.3**
11-PLP DNNs	60.7	63.3	51.1	62.2	59.7	60.0

Table 3: *Effects of adding more data to train BN-DNNs; comparison with intra-lingual BN and 11-PLP baseline.*

BN-DNNs Training Data	Target Language [WER %] Spanish	Mandarin
CH-EN	59.9	62.5
CH-EN+SWB-10h	59.8	62.2
CH-EN+SWB-50h	58.8	61.9
CH-EN+SWB-100h	58.8	61.7
CH-EN+SWB-150h	58.3	61.1
CH-EN+SWB-200h	**58.2**	**61.0**
11-PLP DNNs	60.7	63.3
Intra-lingual BN	59.3	62.0

SWB-10h. **Secondly**, all the BN-DNNs are initialized with the same RBMs which are previously pre-trained on CH-EN data. **Thirdly**, no GMM-HMMs models are newly trained. Labels for each SWB set are obtained by force aligning word transcripts to previously CH-EN trained GMM-HMMs.

The results in Table 3 show that the information extracted by BN-DNNs does increase when adding more training data, though relatively slow. When 50 hours of SWB data is added, cross-lingual BN features have exceeded the intra-lingual features. The best performance is obtained when SWB-200h is added, by 2.5% and 2.3% absolute WER reduction over 11-PLP baseline for Spanish and Mandarin, respectively.

6.4. Concatenation or multilingual BN features?

When multilingual resources available, there are two straightforward approaches to utilize these data under this two-level DNNs framework. The first method is to train mono-lingual BN-DNNs on each language and combine the cross-lingual BN features by use of second-level AM-DNNs. The other method is to apply multi-lingual training with shared hidden layers and language-specific output layers to BN-DNNs. These two approaches are experimentally compared in this section. Spanish and Mandarin are also chosen as two target languages and the results are listed in Table 4. In this work, we focus on cross-lingual BN features, therefore data from target-languages is not used to train BN-DNNs.

The authors in [14] found that multilingual BN features perform slightly better than the concatenation ones. However, the results in this section show that this may not generalize. It can be seen that these two methods, concatenation and multilingual BN features, have similar performance with each other. We attribute the difference with [14] to the different feature configurations (without delta processing and original features in [14]).

Table 4: *Results for two data-utilizing methods: concatenation and multilingual BN features.*

BN-DNNs Multi-languages		Target Language [WER %] Spanish	Mandarin
Concatenate BN Features	EN,GE	59.4	62.1
	EN,GE,AR	58.8	61.6
	EN,GE,AR,JP	58.8	61.2
Multilingual BN Features	EN,GE	59.6	61.9
	EN,GE,AR	59.0	61.4
	EN,GE,AR,JP	58.7	61.3

Another observation which can be obtained from the results is that both of these two methods can increase the recognition performance of this architecture. For example, when English and German resources available, concatenation or multilingual BN features outperform English and German mono-lingual BN features for these two target languages (see Table 2). This observation is slightly different with [10], in which it was found that the combined BN features perform similar to the mono-lingual ones. The reason for this difference may be that the experiments in [10] were situated in tandem framework. The decorrelation and dimensionality reduction steps, which have to be performed before GMM-HMMs training, may harm the resulting combined features. In this two-level DNNs framework, however, the second-level AM-DNNs perform feature transform automatically and have much stronger modeling power than GMMs.

7. Conclusions

This study focuses on the performance of cross-lingual BN features in hybrid ASR systems. BN feature extractors and DNNs acoustic models are combined in a two-level DNNs framework. To obtain the optimal feature configurations, BN features are first optimized experimentally. The results show that delta processing and concatenation with original features are both necessary for BN features, especially for cross-lingual ones. Further experiments are done on all the six languages to evaluate cross-lingual generalization of BN features. It is shown that each of six BN-DNNs has similar feature extraction capabilities. The following experiments analyze the effects of adding more training data on BN-DNNs, showing that the information extracted by BN-DNNs does increase. Finally, we compare two data utilizing methods when multilingual resources available: concatenation and multilingual BN features. The results show that these two approaches have similar performance, and the two-level DNNs architecture benefits from either of them.

8. References

[1] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1635–1638.

[2] F. Grézl, M. Karafiát, S. Kontár, and J. Cernocky, "Probabilistic and bottle-neck features for lvcsr of meetings," in *Proc. ICASSP*, vol. 4, 2007, pp. 757–761.

[3] H. Lin, L. Deng, D. Yu, Y.-f. Gong, A. Acero, and C.-H. Lee, "A study on multilingual acoustic modeling for large vocabulary asr," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4333–4336.

[4] P. N. Garner, D. Imseng, J. Dines, H. Bourlard, and P. Motlicek, "Comparing different acoustic modeling techniques for multilingual boosting," in *Proceedings of Interspeech*, no. EPFL-CONF-192725, 2012.

[5] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4334–4337.

[6] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace gaussian mixture models for cross-lingual speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 365–370.

[7] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8619–8623.

[8] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[9] A. Stolcke, F. Grézl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. 321–324.

[10] F. Grezl, M. Karafiát, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 359–364.

[11] C. Plahl, R. Schluter, and H. Ney, "Cross-lingual portability of chinese and english neural network features for french and german lvcsr," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 371–376.

[12] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 336–341.

[13] J. Gehring, W. Lee, K. Kilgour, I. Lane, Y. Miao, A. Waibel, and S. V. Campus, "Modular combination of deep neural networks for acoustic modeling," in *Proc. Interspeech*, 2013, pp. 94–98.

[14] J. Gehring, Q. B. Nguyen, F. Metze, and A. Waibel, "Dnn acoustic modeling with modular multi-lingual feature extraction networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 344–349.

[15] P. Bell, M. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P. C. Woodland, "Transcription of multi-genre media archives using out-of-domain data," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 324–329.

[16] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid asr systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6975–6979.

[17] H. Christensen, M. Aniol, P. Bell, P. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," 2013.

[18] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks." in *INTERSPEECH*, 2011, pp. 237–240.

[19] F. Grezl and P. Fousek, "Optimizing bottle-neck features for lvcsr," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4729–4732.