INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Acoustic scene classification using teacher-student learning with soft-labels

*Hee-Soo Heo**, *Jee-weon Jung**, *Hye-jin Shim*, *Ha-Jin Yu*†

School of Computer Science, University of Seoul, South Korea

zhasgone@naver.com, jeewon.leo.jung@gmail.com, shimhz6.6@gmail.com, hjyu@uos.ac.kr

Abstract

Acoustic scene classification identifies an input segment into one of the pre-defined classes using spectral information. The spectral information of acoustic scenes may not be mutually exclusive due to common acoustic properties across different classes, such as babble noises included in both airports and shopping malls. However, conventional training procedure based on one-hot labels does not consider the similarities between different acoustic scenes. We exploit teacher-student learning with the purpose to derive soft-labels that consider common acoustic properties among different acoustic scenes. In teacher-student learning, the teacher network produces soft-labels, based on which the student network is trained. We investigate various methods to extract soft-labels that better represent similarities across different scenes. Such attempts include extracting soft-labels from multiple audio segments that are defined as an identical acoustic scene. Experimental results demonstrate the potential of our approach, showing a classification accuracy of 77.36 % on the DCASE 2018 task 1 validation set.

Index Terms: teacher-student learning, knowledge distillation, acoustic scene classification, deep neural networks

1. Introduction

Acoustic scene classification (ASC) refers to a task that categorizes an input audio segment into one of the pre-defined acoustic scenes (classes). The detection and classification of acoustic scenes and events (DCASE) community, the leading platform for ASC and several related tasks, defines 10 acoustic scenes : airport, park, metro station, etc [1]. Such classes possess common acoustic properties, such as babble noises included in both airports and shopping malls.

With advances in deep learning, various deep neural networks (DNNs) are primarily exploited for the ASC task, which are executed in a supervised manner using one-hot labels. However, in the conventional training scheme which use one-hot labels with an softmax output layer, the common properties among different classes cannot be considered. This is because each acoustic scene is trained to be orthogonal in the label space. Here, the term 'orthogonal' refers to having no correlation between any other labels. We hypothesize that this strict training process is not efficient because it is opposite to the process of human perception. Humans can recognize that there is a similarity between different acoustic scenes. This inefficiency of the one-hot label technique can give rise to issues in most identification tasks, which would be further intensified in tasks, such as ASC, where the boundaries or definitions of the scenes are ambiguous. We expected that the similarity of each scene would be reflected in the soft-label.

One of the well-known techniques that extract soft-labels is teacher-student (TS) learning. In the TS learning, the student network is trained using a soft-label from the teacher network rather than conventional one-hot label. The TS framework was first proposed for model compression, but was found to be an effective scheme to deal with various issues in related tasks, such as compensating far-field utterances in speech recognition [2]. The key factor that leads the various applications of TS learning was an appropriate modification of the framework to suit the purpose. Therefore, we exploited various modifications of TS learning and evaluated them to consider common properties among different classes in the ASC task.

With the inspiration from the previous researches [2, 3], we explored two techniques to modify the TS learning to better conduct the ASC task. The first is extraction of soft-labels from multiple input segments. This method enables extracting more general soft-labels and also includes the effect of data augmentation. The second is direct comparison of embeddings rather than the output layer, proposed in [3], was applied for the ASC task. We verified that the combination of the two techniques can further improve the performance of the ASC system.

The rest of the paper is organized as follows. Section 2 describes the overall system, mainly front-end DNN and back-end support vector machine (SVM), used in this study. The TS learning scheme and the proposed techniques are discussed in Section 3. Section 4 presents the experiment and analysis results, and the paper is concluded in Section 5.

2. System description

This section describes the system used for the ASC task. We use a convolutional neural network (CNN) to process the raw waveform, and a CNN with a layer of gated units (CNN-GRU) to process spectrograms. These two front-end models extract a fixed-dimensional embedding from an input segment. We used SVM as the back-end for the ensemble of the two models. The scheme for combining the two models using spectrograms and raw waveforms follows that of the authors' previous research in [4]. In addition, we have improved the performance of the system through few modifications, such as the DNN architecture. Table 1 shows that the baseline systems used in this study outperform the previously reported baselines **without the application of TS learning**.

Table 1: *Classification accuracy (%) of individual systems and their score-sum ensemble in terms of fold 1 configuration on the validation set.*

System	**Jung *et al.*** [1, 4]	**Improved baseline**
Raw waveform	67.15	69.38
Spectrogram	66.24	72.73
i-vector	63.74	-
Ensemble	73.82	**74.42**

*These authors contributed equally

† Corresponding author

http://dx.doi.org/10.21437/Interspeech.2019-1989

Table 2: *DNN architecture of raw waveform model with input sequence shape: (*479999×2*).*

Layer	Output shape	Kernel size	Stride
Conv1	39999 × 64	12	12
Res1	13333 × 64	3	1
Res2	4444 × 128	3	1
Res3	1481 × 128	3	1
Res4	493 × 128	3	1
Res5	164 × 128	3	1
Res6	54 × 128	3	1
Res7	18 × 128	3	1
GlobalPool	128	-	-
Dense1	64	128 × 64	-
Output	10	64 × 10	-

2.1. Extraction of CNN-GRU embedding

State-of-the-art systems in the ASC task comprise deep architectures, using CNNs and recurrent architectures [4–7]. The CNN model for raw waveforms adopts 1D convolutional layers for direct processing of raw waveforms. Utilizing vast audio segments with a high sampling rate, these systems are capable of extracting an embedding that is highly representative.

The DNN used in this study comprises residual convolutional blocks and a fully-connected layer similar to that of [8,9]. In this architecture, the input segments are processed using convolutional layers to extract the frame-level features. These embeddings are then aggregated into an utterance-level feature by using a global max pooling layer. One fully-connected layer is then used to extract the embedding, followed by the output layer. After the training, the output layer is removed, and embeddings are extracted from the last fully-connected layer. Table 2 depicts the overall DNN architecture using raw waveforms.

We used a CNN-GRU model with 2D convolutional layers for processing spectrograms. In this model, a two-channel spectrogram extracted from stereo audio is input to the CNN, and a fixed dimensional embedding is output via the GRU layer. The overall DNN architecture using spectrograms is depicted in Table 3.

Table 3: *DNN architecture of spectrogram model with input sequence shape: (*$249 \times 256 \times 2$*).*

Layer	Output shape	Kernel size	Stride
Conv1	249 × 256 × 30	7 × 7	1 × 1
Res1	249 × 256 × 30	3 × 3	1 × 1
Res2	125 × 128 × 60	3 × 3	2 × 2
Res3	63 × 64 × 120	3 × 3	2 × 2
Res4	21 × 22 × 120	3 × 3	3 × 3
AvgPool	21 × 1 × 240	1 × 22	1 × 22
MaxPool	21 × 1 × 240	1 × 22	1 × 22
Concan	21 × 480	-	-
GRU	480	-	-
Dense1	64	480 × 64	-
Output	10	64 × 10	-

2.2. SVM classification

The DNNs with the output layer activated by the softmax function are well-known as a high-performance classifier. However, the softmax values in the output layer do not represent the concept of confidence. In other words, the softmax values are "poorly calibrated" [10]. In case of a single DNN, this issue does not cause any problems. However, when combining output results from multiple DNNs, this can cause problems because the outputs are not confidence scores. To avoid this problem, we implemented a separate scoring phase using the SVM classifier. In the scoring phase, the output layers of two models are removed and the outputs of the last hidden layer of each model are trained using one SVM for each model. Finally, the scores calculated from the two SVMs are averaged for the ensemble of the raw waveform and spectrogram-based models.

3. Teacher-student learning in ASC

Teacher-student (TS) learning is a framework that adopts two DNNs. In this scheme, a superior system (teacher DNN) is first trained using conventional training scheme with one-hot labels. Superiority of the teacher DNN is determined depending on the target task, i.e., larger capacity for model compression. Then the output of the teacher DNN (referred to as soft-label) is used to train a student DNN [11]. Specifically, the output distributions of the teacher and student DNNs are compared using Kullback-Leibler divergence with the following equation:

$$TS_{org} = -\sum_{i}^{I}\sum_{j}^{J} p_T(o_j|x_i)\log(p_S(o_j|x_i)), \qquad (1)$$

where $p_T(\cdot)$ and $p_S(\cdot)$ are the output distributions of the teacher and student network, respectively, and x_i refers to the $i'th$ input. Eq. (1) refers to cross-entropy rather than KL-divergence, but the same effect can be achieved when training the student network (look [11], Section 3.1. for further details).

The TS framework has been expanded to knowledge distillation with the concept of temperature in [12]. In knowledge distillation, the temperature T adjusts the extent of soft-label utilization where a higher T softens the probability distribution. Hence, the original equation of TS learning is expanded by including temperature variable T as:

$$TS_T = -\sum_{i}^{I}\sum_{j}^{J} p_T(o_j|x_i;T)\log(p_S(o_j|x_i)), \qquad (2)$$

$$p_T(o_j|x_i;T) = \frac{exp(p_T(o_j|x_i)/T)}{\Sigma_k^K exp(p_T(o_k|x_i)/T)}, \qquad (3)$$

3.1. Concatenating multiple inputs

One of the important issues in the TS learning scheme is to design the superiority of the teacher network. In the study that proposed the TS scheme, large capacity of the teacher network was the superiority for a model compression task [11]. For far-field compensation study, less noise and reverberation comprised the superiority where the utterances recorded from close talk and the utterances recorded from far-field are input to the teacher and the student respectively [2, 14]. In our study, possession of additional recording from identical acoustic scenes is set as the superiority. For this superiority, additional recording from the same class is concatenated to every utterance, and then used for extracting the soft-label. For instance, the soft-label of a student DNN with input utterance 'A' will be derived by concatenating another utterance 'B' from the same class and then inputting to the teacher DNN. Here, utterance 'A' is input to the student DNN and concatenation of utterance 'A', 'B' is input to

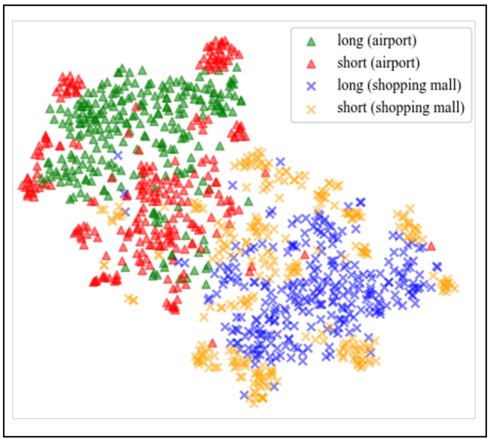

Figure 1: *Illustration of the superiority using multiple segments for longer duration for the teacher network. t-SNE [13] plot of embeddings extracted from test set segments on the most confusing acoustic scene pair, 'shopping_mall-airport'.* $\triangle$ *and* $\times$ *symbols represent the embeddings of shopping_mall and airport scene, respectively. Cohesion of long duration embeddings are stronger, demonstrating the superiority of teacher network's input.*

the teacher DNN to extract a soft-label. This scheme is represented using the following equation:

$$KL_{loss} = -\sum_{i}^{I}\sum_{j}^{J} p_T(o_j|x_{i,con})\log(p_S(o_j|x_{i,base})), \quad (4)$$

where $x_{i,base}$ is the base segment and the input segment of the student network, and $x_{i,con}$ is the input segment of the teacher network constructed by concatenating the input segment of the student DNN with another segment from an identical class. In particular, the length of $x_{i,base}$ is fixed to 10 s, and the duration of $x_{i,con}$ can be 20, 30 s, or longer. The superiority caused by concatenating multiple inputs is shown in detail in figure 1. The figure shows that the embeddings extracted from long segments have higher discriminative power. This superiority is also interpreted as that of short utterance compensation in the field of speaker verification [3, 15].

Deriving soft-labels using multiple audio recordings is also expected to include the effect of data augmentation. Training data augmentation, conducted by adding noises or shifting pitch, is a well-known technique for performance improvements in the audio processing tasks using DNNs. In the ASC task, however, it is difficult to apply the conventional data augmentation scheme because the definition of acoustic scene is ambiguous, and there is no clear approach to define which noise belongs to which class. In particular, adding babble noise to the training data can give rise to a critical issue that changes the class label of the training data. In the proposed approach, the input data of the teacher network is changed depending on the selection of inputs for concatenation where the input audio recording of the student network is fixed. Therefore, we expected that there would be a similar effect of data augmentation through these combinations of multiple inputs.

3.2. Learning based on embeddings distance

In the proposed ASC system introduced in Section 2, the softmax output layer is removed after the training phase and the embeddings extracted from the last hidden layer are used for the scoring phase. Therefore, the last hidden layer, and not the output layer, is a more dominant factor in the performance of the ASC system. Based on this property, we modified the TS learning to take into account the output of the last hidden layer as follows:

$$TS_{emb} = \sum_{j}^{J} Dist(E_T(x_j), E_S(x_j)), \quad (5)$$

where $Dist(\cdot)$ is the distance measure between two vectors, and $E_T(\cdot)$ and $E_S(\cdot)$ are the output of the last hidden layer from the teacher and the student network, respectively. In this study, we used mean squared error as the distance measure. We interpreted this approach as the distilling the knowledge at the last hidden layer, and not at the output layer. This approach was inspired by [3] and [2].

4. Experiments and results

In this section, we show the results of experiments to evaluate the various TS learning techniques. The baseline system used for the performance comparison is an improved version of the authors' system that was presented at the last DCASE 2018 competition (see Table 1). Therefore, we focused on evaluating the effectiveness of the TS learning rather than comparing it with other systems.

4.1. Dataset

DCASE 2018 task 1-a dataset [1] was used for all experiments in this study. This dataset comprises 864 audio segments that has 10 s duration for each of 10 pre-defined classes, resulting in a total of 8640 segments. The audio segments were recorded in stereo at a sampling rate of 48 kHz. Cross-validation was conducted using the four fold configuration provided within the DCASE dataset, where the validation sets are recorded from different locations.

4.2. Experimental settings

All experiments in this study were conducted using Keras, a deep learning library for python, with Tensorflow back-end [16–18].

For the raw waveform model, pre-emphasis is applied [19] without any other pre-processing. The whole segment is input, which makes the shape of input segment as $(479999, 2)$. The DNN architecture configuration is depicted in Table 2.

We extracted the spectrograms of 256 coefficients from 100 ms windows for every 40 ms. The spectrogram of $(249, 256, 2)$ shape was extracted for each segment that contains stereo audio of 10 s.

Adam optimizer [20] with 0.001 learning rate was used for training both CNN using raw waveforms and CNN-GRU using spectrograms. The batch size for training the two models was 40. For efficient training of the spectrogram-based CNN-GRU model, we trained the CNN part except the GRU layer in the whole model and then re-trained after attaching the GRU layer on the CNN, following the multi-step training scheme reported in [9, 21].

4.3. Analysis of results

Table 1 demonstrates the baseline of this study, which is an improved version of the authors' submission to DCASE 2018

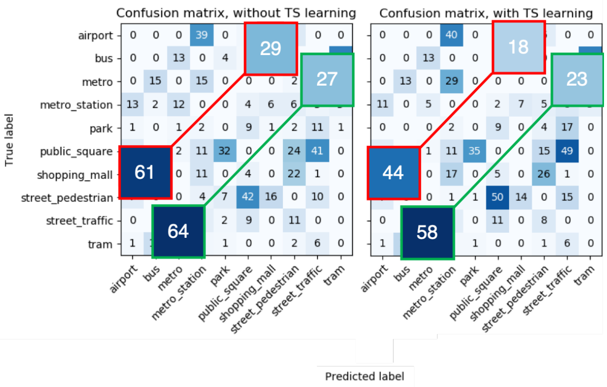

Figure 2: *Confusion matrices before (left) and after (right) applying the proposed TS learning scheme. Both of the two most confusing acoustic scene pairs, 'shopping_mall-airport' (red box) and 'metro-tram' (green box) show improvement.*

competition. The individual systems show improved performance. For the ensemble, "Improved baseline" outperforms the previous baseline despite the i-vector [22] system is excluded.

Table 4 shows the performances of spectrogram-based models using TS learning with various temperature T and input durations. The temperature coefficient T, described in equation (2), is an important factor that determines the extent of soft-label utilization. As the value of T increases, the output distribution of the teacher network becomes more noisy. On the contrary, decrease in the value of T refers to imposing more weight to the one-hot (hard) label. Comparing the results when T values is one and five, we interpret that to some extent, considering common properties is important.

It is also important to design the superiority of the teacher network in the TS learning scheme. This is because the training process of the student network is totally dependent on the teacher network. However, the teacher network, trained using one-hot labels, may have poor performance. Even the teacher network may distill wrong knowledge to the student network. Experimental results showed that adjusting the input length of the teacher network to gain superiority could lead to performance improvements as intended. Table 5 demonstrates the effectiveness of direct comparison of the embeddings between the teacher and student networks. The results show that use of the embedding of teacher DNN outperforms the soft-label of the output layer. Use of both soft-label and embedding layer, however, shows decreased accuracy. We interpret that this phenomenon occurred because the common properties across different classes are better represented in the embedding space rather than using human defined one-hot labels. For example, the soft-label generated at the output layer represents the common properties, such as babble noise depending on the corresponding classes (airport or shopping mall), but the soft-label at the last hidden layer can represent the common property by manifold in a high dimensional embedding space. Table 6 shows the comparison of the approaches used in this study to the baseline. The proposed approach with TS learning demonstrates classification accuracy of 77.36 % compared to 74.42 % for the scheme without TS learning. From these results, we conclude that the TS learning scheme is effective for the ASC task, and the approaches developed in this study are also valid. We analyzed the detailed contribution of performance enhancement by TS learning. Figure 2 illustrates two confusion matrices, with and without applying the proposed TS learning scheme, which shows that TS learning significantly reduced the errors between the two most confusing acoustic scene pairs.

Table 4: *Performance in terms of accuracy (%) depending on temperature T and superiority of the teacher network.*

		Duration of teacher input	
		10 seconds	20 seconds
T	1	70.96	71.43
	5	72.63	**73.23**
	10	72.51	72.79

Table 5: *Performance depending on points of knowledge distillation.*

Point	accuracy (%)
output layer	73.23
output & last hidden layer	73.19
last hidden layer	**74.26**

Table 6: *Classification accuracy of individual systems and their score-sum ensemble in terms of fold 1 configuration on the validation set (W/O TS: systems trained without TS learning, W/ TS: systems trained with TS learning).*

System	**W/O TS**	**W/ TS**
Raw waveform	69.38	72.99
Spectrogram	72.59	74.26
Ensemble	74.42	**77.36**

5. Conclusions and discussion

The conventional training procedures using one-hot label cannot represent common properties among different classes. We assume that this scheme is not appropriate for tasks such as the ASC where the decision boundary of each class is ambiguous. Therefore, we explored various applications of TS learning to use the soft-label instead of the one-hot label. We applied TS learning to the ASC task for the first time based on two techniques: using multiple segments for the teacher network and distilling the knowledge at the last hidden layer. In TS learning, the student network is trained using soft-labels extracted from the teacher network. Soft-label in TS learning was interpreted to incorporate the correlation of different acoustic scenes with common acoustic properties. We evaluated various approaches of TS learning using the DCASE 2018 task 1 dataset. Experimental results demonstrate that designing the superiority of the teacher network and adjusting the point of knowledge distillation could improve the performance. In particular, TS learning significantly reduced the errors between the most confusing scenes.

6. Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(2017R1A2B4011609)

7. References

[1] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," 2018, submitted to DCASE2018 Workshop. [Online]. Available: https://arxiv.org/abs/1807.09840

[2] J. Kim, M. El-Khamy, and J. Lee, "Bridgenets: Student-teacher transfer learning based on recursive neural networks and its application to distant speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5719–5723.

[3] J.-w. Jung, H.-s. Heo, H.-j. Shim, and H.-j. Yu, "Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings," *arXiv preprint arXiv:1810.10884*, 2018.

[4] ——, "DNN based multi-level feature ensemble for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 113–117.

[5] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," DCASE2018 Challenge, Tech. Rep., September 2018.

[6] M. Dorfer, B. Lehner, H. Eghbal-zadeh, H. Christop, P. Fabian, and W. Gerhard, "Acoustic scene classification with fully convolutional neural networks and I-vectors," DCASE2018 Challenge, Tech. Rep., September 2018.

[7] H. Zeinali, L. Burget, and H. Cernocky, "Convolutional neural networks and x-vector embedding for dcase2018 acoustic scene classification challenge," DCASE2018 Challenge, Tech. Rep., September 2018.

[8] J. Jung, H. Heo, I. Yang, H. Shim, and H. Yu, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5349–5353.

[9] ——, "Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification," in *Proc. Interspeech 2018*, 2018, pp. 3583–3587.

[10] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *arXiv preprint arXiv:1706.04599*, 2017.

[11] J. Li, R. Zhao, J. Huang, and Y. Gong, "Learning small-size dnn with output-distribution-based criteria," in *Fifteenth annual conference of the international speech communication association*, 2014.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[13] L. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, 2008.

[14] J. Li, R. Zhao, Z. Chen, C. Liu, X. Xiao, G. Ye, and Y. Gong, "Developing far-field speaker system via teacher-student learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5699–5703.

[15] H. Yamamoto and T. Koshinaka, "Denoising autoencoder-based speaker feature restoration for utterances of short duration," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.

[17] A. Martín, A. Ashish, B. Paul, B. Eugene *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2015. [Online]. Available: http://download.tensorflow.org/paper/whitepaper2015.pdf

[18] A. Martin, B. Paul, C. Jianmin, C. Zhifeng, D. Andy, D. Jeffrey, D. Matthieu, G. Sanjay, I. Geoffrey, I. Michael, K. Manjunath, L. Josh, M. Rajat, M. Sherry, M. G. Derek, S. Benoit, T. Paul, V. Vijay, W. Pete, W. Martin, Y. Yuan, and Z. Xiaoqiang, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283. [Online]. Available: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[19] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition," in *Electrical and Computer Engineering, 1995. Canadian Conference on*, vol. 2. IEEE, 1995, pp. 1062–1065.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] H. S. Heo, J. W. Jung, I. H. Yang, S. H. Yoon, and H. J. Yu, "Joint training of expanded end-to-end DNN for text-dependent speaker verification," *Proc. Interspeech 2017*, pp. 1532–1536, 2017.

[22] 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*, vol. 19, no. 4, pp. 788–798, 2011.