RECOGNITION OF ACOUSTIC EVENTS USING DEEP NEURAL NETWORKS

Oguzhan Gencoglu, Tuomas Virtanen, Heikki Huttunen

Department of Signal Processing, Tampere University of Technology, 33720 Tampere, Finland

ABSTRACT

This paper proposes the use of a deep neural network for the recognition of isolated acoustic events such as footsteps, baby crying, motorcycle, rain etc. For an acoustic event classification task containing 61 distinct classes, classification accuracy of the neural network classifier (60.3%) excels that of the conventional Gaussian mixture model based hidden Markov model classifier (54.8%). In addition, an unsupervised layer-wise pretraining followed by standard backpropagation training of a deep network (known as a deep belief network) results in further increase of 2-4% in classification accuracy. Effects of implementation parameters such as types of features and number of adjacent frames as additional features are found to be significant on classification accuracy.

Index Terms**—** acoustic event classification, artificial neural networks, deep belief networks, deep neural networks, pattern classification.

1. INTRODUCTION

Acoustic event recognition addresses the recognition of sounds which are generated by nature, by objects handled by humans or by humans themselves. Classification and detection of these sounds, namely acoustic events, is useful in information retrieval, having applications in multimedia content analysis, context-aware devices and audio-based surveillance and monitoring systems.

Some of the research on acoustic event classification has focused on classification of acoustic events into event classes for a specific context [1, 2]. Some other has focused on classification of acoustic events into contextual classes [3, 4]. Throughout these works, varying classification rates have been achieved depending on the complexity of the problem (number of different classes, available number of data, quality of the data, distribution of the data etc.) The features used to represent the audio data and the classifiers used for the classification task also differ from work to work.

In [5], a k-nearest neighbor classifier was established for classification of certain acoustic events. Automatic speech recognition algorithms such as Gaussian mixture models (GMM) with hidden Markov models (HMM) [3, 4, 6, 7] have been the most commonly used methods. Other methods such as vector quantization [8], decision trees [9] and support vector machines [10, 11] have also been tried. Among these methods, GMM based HMM was considered to be the standard classifier in acoustic event classification. In [6], acoustic events corresponding to 61 different classes were classified with such a classifier, achieving a classification accuracy of 54.8%.

Classification accuracies of the abovementioned classifiers for acoustic event classification tasks point out presence of room for improvement. Thus, search for enhancement of these methods or investigation of new approaches is essential. With the help of new training mechanisms, deep neural networks (DNN) are giving promising results in many pattern recognition applications. Recently, in automatic speech recognition, DNNs have outperformed the conventional approaches [12] and due to the similar nature of the problems DNNs are worth to investigate for acoustic event classification tasks.

In our work, a deep neural network classifier is proposed to perform acoustic event classification. In addition, the power of neural network (NN) classifiers is underlined as well as the advantage of unsupervised pre-training on the DNN performance for a given acoustic event classification task. Our approach shows significant improvement of classification performance over other methods for acoustic event classification tasks including the standard GMM based HMM classifier.

This paper is organized as follows. Section 2 presents the task of acoustic event classification using deep neural networks. Evaluation setup is presented in Section 3. In Section 4, main results of conducted experiments as well as the effect of certain network and implementation parameters on classification accuracy are presented. Finally, in Section 5, conclusions are drawn.

2. ACOUSTIC EVENT CLASSIFICATION USING DEEP NEURAL NETWORKS

The schematic of the proposed acoustic event classification system can be seen in Figure 1. Firstly, input audio files are preprocessed with amplitude normalization, frame blocking and windowing. After that, feature extraction is applied to each frame separately in order to represent the audio data by a set of acoustic features. Due to the presence of silent frames, a fixed number of most energetic frames are selected from each audio file by discarding the rest. The number of frames

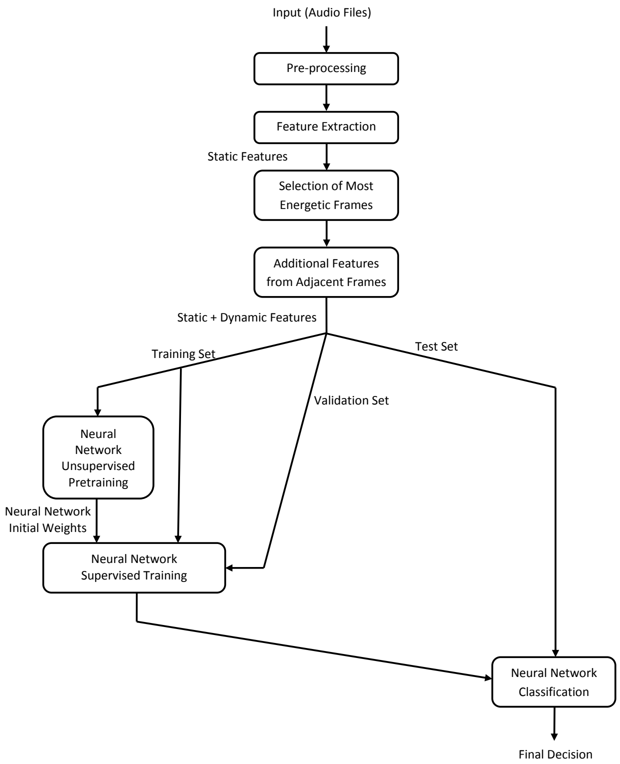
Input (Audio Files)
Pre-processing
Feature Extraction
Static Features
Selection of Most Energetic Frames
Additional Features from Adjacent Frames
Static + Dynamic Features
Training Set
Test Set
Validation Set
Neural Network Unsupervised Pretraining
Neural Network Initial Weights
Neural Network Supervised Training
Neural Network Classification
Final Decision

Fig. 1. Acoustic event classification system schematic.

coming from each file is also kept the same with this approach independent of the file length. To model the dynamic properties of sounds, adjacent frames are also taken into consideration. The number of features that represent each frame is increased by concatenation of features of the current frame with that of left and right adjacent frames. The data is then divided into three distinct sets namely, training, validation and test set. Training set is used to train the neural network classifier first in an unsupervised manner. Then a supervised training follows that, which is conducted simply by introducing labeled examples to the network. Validation set is used to tune the neural network training parameters and to adjust the neural network topology. It has a significant role in the decision of stopping the supervised neural network training as well. Finally, test set is simply used for the performance evaluation of the trained neural network classifier.

2.1. Deep neural networks

Neural networks are powerful pattern classifiers which have been used in numerous classification and function approximation tasks. They are highly nonlinear classifiers not only because they have nonlinear activation units but also because of the layer-wise structure stacked one after another. Such a structure enables the NNs to learn the complex input-output relationships of many classification problems such as acoustic event classification.

Artificial neural networks are trained in a supervised manner with the backpropagation algorithm in which the randomly initialized network weights are adjusted according to the gradient descent rule to learn the input-output relations from labeled data. Backpropagation algorithm performs effectively for shallow networks, i.e., those that have 1 or 2 hidden layers, but its performance declines when the number of layers increases. Numerous experiments show that the algorithm gets stuck in local optima easily and fails to generalize properly for deep networks [13, 14] (with a possible exception of convolutional neural networks, which were found to be easier to train even for deeper architectures [15, 16]). In general, it is shown that, when NN weights are randomly initialized, deep neural networks perform worse than the shallow ones [13, 17].

In order to ease the training of deep networks, an unsupervised pre-training is conducted layer by layer, to initialize the network weights [18]. This greedy, layer-wise unsupervised pretraining is based on restricted Boltzmann machine (RBM) generative model. An algorithm called contrastive divergence (CD) is applied to train an RBM. CD algorithm trains the first layer in an unsupervised manner, producing an initial set of coefficients for the first layer of a NN. Then, the output of the first layer is fed as an input to the next, again initializing the corresponding layer in an unsupervised way and so forth. The mathematical details of the CD algorithm, can be found in [19] and will not be presented in this work. After pretraining, neural networks are trained in a supervised manner with batch backpropagation algorithm in which the weight updates take place after a number of training samples is presented to the network (batch size). This step serves as a fine-tuning process of the neural network coefficients which were initialized with pretraining.

In this work, the topology of the neural network (5 hidden layers each containing 70 neural units with sigmoid activation functions) is chosen according to the validation set and the effect of variations in network topology on the classification accuracy is not presented. Training parameters for the neural networks such as learning rate, momentum, batch size etc. as well as their topology are kept the same for all of the experiments. The batch size for both unsupervised and supervised parts are chosen to be 100. Learning rate and momentum of backpropagation are decided to be 0.5 and 1 respectively. The number of epochs for the unsupervised pretraining is fixed to 2. The criterion for stopping the supervised training was based on the validation set error. The training was terminated when the validation error started to increase which is a sign of overfitting.

2.2. File classification

In our work, classification of audio files belonging to distinct classes is aimed. Each WAVE file is normalized to unity maximum amplitude and blocked into frames of 50 ms with Hamming windowing function with 50% overlap in the preprocessing step. 40 mel energy coefficients are extracted from each frame. For each file, 120 most energetic frames are taken into consideration. Additional features are provided by adjacent frames of these most energetic ones. The number of left and right adjacent frames for providing extra features is 2. Therefore, each input observation to the network training consists of 200 coefficients (40 from the current frame, 80 left and 80 right adjacent) and there are 120 observations for each file. During testing, each frame outputs a set of values which can be considered as likelihoods. A test file is classified to the class that gives the maximum value for the mean of outputs of 120 frames belonging to that file.

Effect of different representations of audio data on the classification accuracy is also examined. The investigated features are MFCCs, mel energies and log-mel energies. Effect of number of adjacent frames for additional features on the classification performance is one other thing that is investigated. The number of adjacent frames is varied from 0 to 7 for that purpose.

3. EVALUATION SETUP

The database used in all of the experiments, which was first introduced in [6], includes acoustic events of 61 distinct classes such as sneezing, dog barking, clapping, car door, beep, yelling. The database is a collection of isolated sound events which was retrieved from the Stockmusic online sample database. There are in total 1325 audio files. Both the number of files per class and the average length of files per class is very uneven which makes the classification problem rather difficult. The minimum number of files for a class is 10. The maximum number of files for a class is 94. The shortest file is about 0.3 milliseconds and the longest file is about 3 minutes 46 seconds long. The histogram of number of files per class and the histogram of average number of frames per class can be seen in Figure 2 and Figure 3, respectively. The second histogram was plotted assuming a frame length of 50 ms to give a comparison of average length of files per class. The database is highly heterogeneous both in terms of number of files per class as well as length of files per class. Together with the high number of classes, such a classification task can be considered rather difficult.

In our experiments the audio files are randomly divided into training, validation and test data as 80%, 10% and 10% respectively at all times. A 10-fold cross-validation is performed and the classification rates are the averaged rates of the all 10 folds. All classification accuracies in this work are calculated according to the number of files (not to the number of frames) that are correctly classified.

The implementation platform for the whole work was MATLAB. For neural network implementations, an open source toolbox, i.e., DeepLearnToolbox [20] was used.

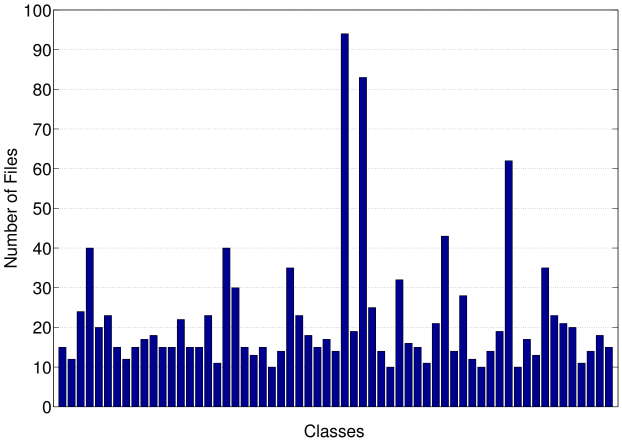

Fig. 2. Histogram of number of files per class.

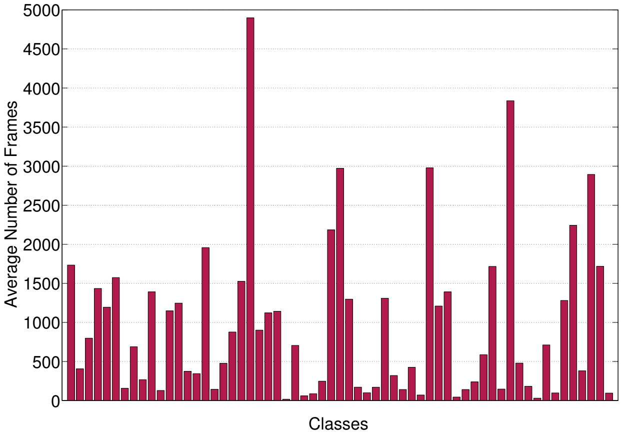

Fig. 3. Histogram of average length of files per class.

The baseline classifier is based on [6], i.e., a three-state, left-to-right HMM with 8 mixture densities. Static and dynamic mel energies were used as features and HMMs were trained using the Expectation-Maximization algorithm. Viterbi algorithm was established in the classification stage. With the validation and training sets combined, same 10-fold cross-validation yields a classification rate of 54.8%.

4. EXPERIMENTAL RESULTS

The classification accuracies obtained by different recognition systems are given in Table 1. The main result of this work is that the proposed pretrained NN classifier with 5 hidden layers outperforms both the baseline GMM + HMM classifier and the 2 layer NN classifier. When mel energy features are used with 2 left and right adjacent frames for additional features, pretrained NN classifier achieves a classification accuracy of 64.6% (class-wise accuracies in Figure 4). Accuracy of the baseline and 2-layer NN classifiers on the same database are 54.8% and 60.2% respectively. Note that, as expected, without pretraining a rather deep network topology fails to perform as good as a shallow one.

Classifier	Classification Accuracy (%)
GMM + HMM	54.8
NN (2 layers)	60.2
NN (5 layers)	38.7
NN (5 layers - pretrained)	**64.6**

Table 1. Classification accuracies for different types of classifiers and topologies.

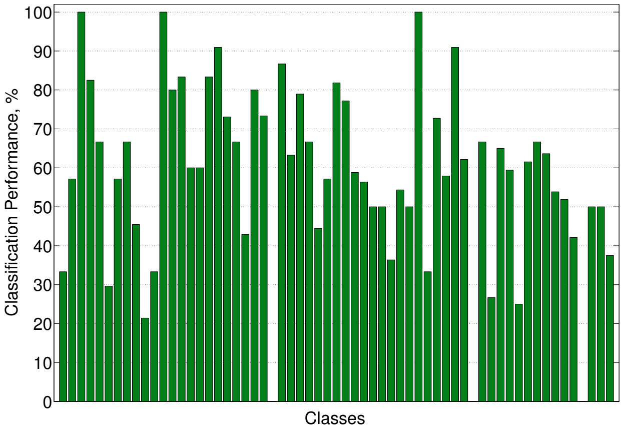

Fig. 4. Classification accuracies for each class.

4.1. Effect of unsupervised pretraining

The effect of unsupervised pretraining is examined for different number of neural network depths (Figure 5). The classification accuracy of 60.2% is achieved for a 2-hidden-layer NN. The classification performance decreases gradually as the depth (number of layers) of the network increases as backpropagation algorithm gets stuck in local optima. The unsupervised pretraining provides better initializations to the network weights, resulting in 64.6% classification rate at 5 layers. Note that the effect of unsupervised, greedy, layer-wise pretraining can not be observed for shallow networks.

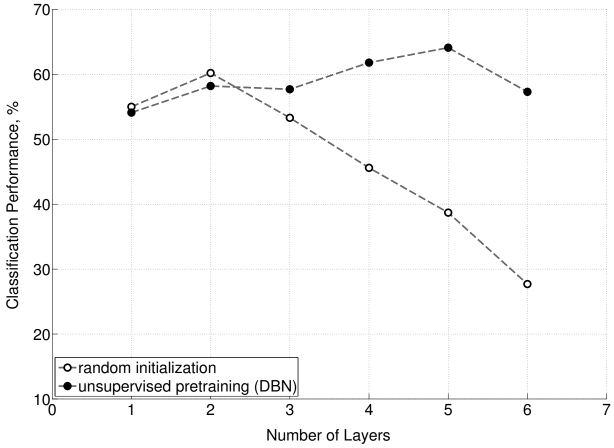

Fig. 5. Effect of unsupervised pretraining on classification accuracy for different number of hidden layers.

Features	Classification Accuracy (%)
MFCCs	53.3
mel energies	64.6
log-mel energies	63.2

Table 2. Classification accuracies for different types of features.

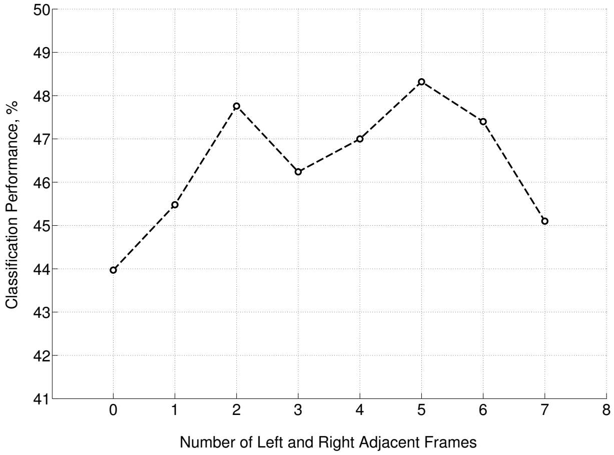

Fig. 6. Effect of number of adjacent frames on classification accuracy.

4.2. Effect of type of features

The effect of type of features is examined on the pretrained DNN structure of 5 hidden layers with 2 adjacent frames. The results are presented in Table 2. The MFCCs give a classification accuracy of 53.3%. Mel energies and log-mel energies achieve classification accuracies of 64.1% and 63.2% respectively.

4.3. Effect of number of adjacent frames for additional features

The adjacent frames were used to provide extra features to represent the dynamic properties of sounds. The effect of number of left and right adjacent frames on the classification accuracy of a 2 layer NN can be seen from Figure 6. MFCCs were chosen to be the features in this case. One can conclude that small number of adjacent frames fail to represent the dynamic properties well enough. On the other hand, too large number of adjacent frames result in a decrease in performance too. This may be due to encountering of silent frames.

5. CONCLUSIONS

In this work, the classification performance of deep neural networks is shown to be considerably better than that of conventional audio classifiers that utilizes HMMs with GMMs on the same database [6]. Classification accuracy of 64.6% is reached on the audio file database of 1325 files belonging to 61 isolated event classes. In addition, unsupervised pretraining is found to be beneficial in terms of classification accuracy. Mel energies are found to give a better represent of the acoustic data compared to other features, i.e., MFCCs or log-mel energies. The effect of retrieving additional features from adjacent frames is also found to be significant on classification accuracy.

For future investigation, the pretrained DNN classifier could be evaluated in more complex acoustic event classification scenarios, i.e., more number of classes, noisy recordings etc. Multi-label classification of audio files containing more than one acoustic events could also be investigated.

In addition, the effect of network topology (number of layers, number of units in each layer, type of activation function) and training parameters (batch size, learning rate, momentum etc.) on the classification accuracy could be examined in the future. Under better circumstances, the benefit of unsupervised pretraining could be more apparent.

REFERENCES

[1] C.-H. Lee, C.-C. Han, and C.-C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1541–1550, Nov 2008.

[2] P. Herrera-Boyer, G. Peeters, and S. Dubnov, "Automatic classification of musical instrument sounds," *Journal of New Music Research*, vol. 32, no. 1, pp. 3–21, 2003.

[3] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 321–329, Jan 2006.

[4] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–13, 2013.

[5] M. Stager, P. Lukowicz, N. Perera, T. von Buren, G. Troster, and T. Starner, "Soundbutton: design of a low power wearable audio classification system," in *Wearable Computers, 2003. Proceedings. Seventh IEEE International Symposium on*, Oct 2003, pp. 12–17.

[6] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *18th European Signal Processing Conference*, 2010, pp. 1267–1271.

[7] M. Slaney, "Mixtures of probability experts for audio retrieval and indexing," in *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, 2002, vol. 1, pp. 345–348 vol.1.

[8] M. Cowling and R. Sitte, "Recognition of environmental sounds using speech recognition techniques," in *Advanced signal processing for communication systems*, pp. 31–46. Springer, 2002.

[9] D. Hoiem, Y. Ke, and R. Sukthankar, "Solar: Sound object localization and retrieval in complex audio environments," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. IEEE, 2005, vol. 5, pp. v–429.

[10] L. S. Kennedy and D. P. W. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal*. National Institute of Standards and Technology, 2004, pp. 118–121.

[11] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *Neural Networks, IEEE Transactions on*, vol. 14, no. 1, pp. 209–215, 2003.

[12] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[13] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagating errors*, MIT Press, Cambridge, MA, USA, 1988.

[14] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[15] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 609–616.

[16] P. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis.," in *ICDAR*, 2003, vol. 3, pp. 958–962.

[17] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The difficulty of training deep architectures and the effect of unsupervised pre-training," in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 153–160.

[18] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[19] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[20] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," M.S. thesis, Technical University of Denmark, 2012.