

A Study on Far-field Emotion Recognition Based on Deep Convolutional Neural Networks

Panikos Heracleous¹, Yasser Mohammad^{1,2}, Koichi Takai¹, Keiji Yasuda¹,
Akio Yoneyama¹, Fumiaki Sugaya^{1,3}

¹KDDI Research, Inc., Japan

2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502, Japan

{pa-heracleous, ko-takai, ke-yasuda, yoneyama}@kddi-research.jp

²Artificial Intelligence Research Center, AIST, Japan

yasserm@aun.edu.eg

³MINDWORD Inc.

7-19-11 Nishishinjuku, Shinjuku-ku, Tokyo, 160-0023, Japan

fsugaya@mindword.jp

Abstract. Automatic recognition of human emotions is a relatively new field, and is attracting significant attention in research and development areas because of the major contribution it could make to real applications. The current study focuses on far-field speech emotion recognition using the state-of-the-art spontaneous IEMOCAP emotional data. For classification, a method based on deep convolutional neural networks (DCNN) and extremely randomized trees is proposed. The method is also compared to support vector machines (SVM) and probabilistic linear discriminant analysis (PLDA) classifiers in the i-vector paradigm. When reverberant speech was classified using the proposed method, the classification rates were comparable to those obtained when using clean data. In the case of PLDA and SVM classifiers, the classification rates were significantly decreased. To further improve the performance of far-field speech emotion recognition, a method based on multi-style training is proposed, which results in significant improvements in the classification rates.

Keywords: Speech emotion recognition, far-field, deep convolutional neural networks, i-vectors, multi-style training approach.

1 Introduction

Emotion recognition plays an important role in human-machine communication [4]. Emotion recognition can be used in human-robot communication, when robots communicate with humans in accord with the detected human emotions, and also has an important role in call centers to detect the caller's emotional state in cases of emergency (e.g., hospitals, police stations), or to identify the level of the customer's satisfaction (i.e., providing feedback). In the current study, emotion recognition based on speech in clean, noisy, and reverberant environments is experimentally investigated.

Previous studies reported automatic speech emotion recognition using Gaussian mixture models (GMMs) [34], hidden Markov models (HMM) [30], support vector machines (SVM) [22], neural networks (NN) [21], and deep neural networks (DNN) [31]. In [17, 38], speech emotion recognition using i-vector features and SVM is described. In [37], a study based on concatenated i-vectors is reported. Audiovisual emotion recognition is presented in [18]. The majority of studies dealing with speech emotion recognition address the problem using acted speech recorded with a close-talking microphone in a clean environment. In fact, only a few studies have used spontaneous or elicited emotion data. For real-world application, however, noise and reverberation in speech emotion recognition must also be addressed and analyzed. Previous studies [10, 29] have investigated the noise issue in speech emotion recognition by using data with superimposed white noise. In [35], several kinds of noise were recorded and superimposed onto clean data to simulate noisy emotional speech data, and adaptive noise cancellation was used as front-end to speech emotion recognizer. In [9], robust speech emotion recognition using a denoising autoencoder was used. In [23], spectral and cepstral audio denoising techniques were applied in speech emotion recognition.

The current study focuses on far-field speech emotion recognition, when the speaker is assumed to be located far way from the microphone and the speech emotion recognition system is being used in a hands-free mode. In such cases, the speech signal is also contaminated with environmental noise and reverberations. In the current study, the reverberant environments were simulated using real impulse responses recorded in rooms of different reverberation times, and convoluted with clean speech. Furthermore, real room noise was superimposed onto the reverberant data to simulate a more realistic situation. Instead of acted emotional speech, the state-of-the-art spontaneous emotional speech database IEMOCAP [3] was used. For classification, deep convolutional neural networks (DCNN) [1, 15] along with extremely randomized trees [8], multi-class SVM [5], and probabilistic linear discriminant analysis (PLDA) [13] classifiers were used. In the case of SVM and PLDA, i-vectors extracted from spectral features were used. Due to the limited amount of training data, i-vectors were not applied in the case of using DCNN. Instead, mel-frequency cepstral coefficients (MFCCs) [28] along with shifted delta cepstra (SDC) coefficients [2, 33] were used to extract informative features. The extracted features were then used by extremely randomized trees for emotion classification.

To improve robustness against noise and reverberation, a method based on multi-style training [24] is proposed. The authors were interested in whether i-vectors can capture the multiple noise and reverberation variability in the case of multi-style training, and how extremely randomized trees, SVM, PLDA with multi-style training perform in the case of far-field speech emotion recognition. Multi-style training is widely used in automatic speech recognition, and in the current study is adapted to speech emotion recognition. Specifically, the training data consist of data for different reverberation times, and do not include the same reverberation as the test data (i.e. reverberation-independent). In the current study, the proposed multi-style training is applied in all stages of the i-vector

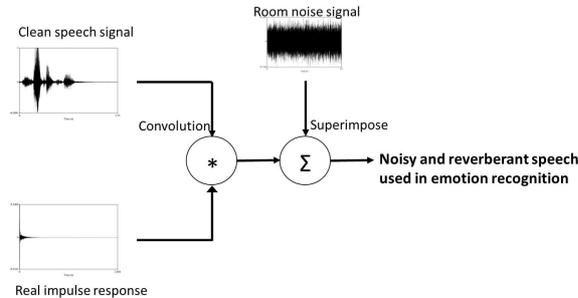


Fig. 1: Producing noisy and reverberant data using real impulse responses and real room noise.

extraction, and not only for training the universal background model (UBM) and to compute the T total variability matrix.

2 Methods

2.1 Data

In the current study, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) spontaneous emotional databases is used. The IEMOCAP database is an acted, multimodal and multispeaker database, collected at the SAIL lab of the University of Southern California, and it contains 12 hours of audiovisual data produced by ten actors. Specifically, the IEMOCAP database includes video, speech, motion capture of face, and text transcriptions. It consists of dyadic sessions where actors perform improvisations or scripted scenarios, specifically selected to elicit emotional expressions. The IEMOCAP database is annotated by multiple annotators into several categorical labels, such as anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation and dominance. In the current study, categorical labels are being used to classify the emotional states of neutral, happiness, anger, and sadness. To avoid unbalanced data, for training, 250 utterances and for testing 70 utterances randomly selected from each emotion were used.

2.2 Reverberant and Noisy Data

The reverberant data are produced using impulse responses convoluted with the clean data. Instead of simulated impulse responses (e.g., using the image method), in the current study real impulse responses recorded in five rooms with different $T_{[60]}$ reverberation times are being used. For recording, a linear microphone array with 14 transducers located at 2.83cm intervals is used [20]. The impulse response is measured using the TSP method [32]. TSP length is 65536 points. The number of synchronous additions is 16. Impulse responses in

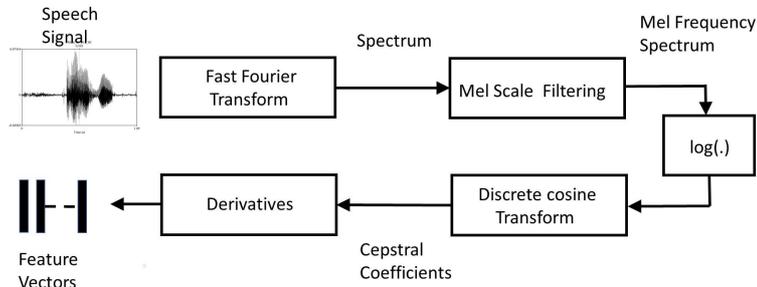


Fig. 2: Extraction of Mel-frequency Cepstral Coefficients (MFCC).

five different rooms are being recorded. The $T_{[60]}$ reverberation times are 0.30, 0.47, 0.60, 0.78, and 1.3 seconds, respectively. The reverberant data are obtained using a convolution method to convolute the clean data with the impulse responses. Furthermore, real room noise (i.e., kitchen fan) at 20dB signal-to-noise ratio (SNR) level is superimposed on the reverberant data to simulate a realistic noisy and reverberant speech emotion recognition environment. The reverberant data are produced using the following formula:

$$y(t) = x(t) * h(t) + n(t) \quad (1)$$

where $y(t)$ is the noisy and reverberant data, $x(t)$ is the close-talking speech data, $h(t)$ is the impulse response, and $n(t)$ is the additive noise. Figure 1 shows the method used to produce noisy and reverberant data.

2.3 Feature selection

Mel-frequency cepstral coefficients (MFCCs) [28] are used in the experiments. MFCCs are very popular features in speech recognition, speaker recognition, emotion recognition, and language identification. Specifically, in the current study, 12 MFCCs plus energy are extracted each 10 ms using a window-length of 20 ms. Figure 2 shows the block diagram of MFCC extraction.

Shifted delta cepstral (SDC) coefficients have been successfully used in language recognition. In the current study, the use of SDC features in speech emotion recognition is also experimentally investigated in order to increase the temporal information in the feature vectors. The SDC features are obtained by concatenating delta cepstral across multiple frames. In this study, the SDC features are also used for speech emotion recognition, along with the MFCC features. The SDC features are described by four parameters, N , d , P and k , where N is the number of cepstral coefficients computed at each frame, d represents the time advance and delay for the delta computation, k is the number of blocks whose delta coefficients are concatenated to form the final feature vector, and P is the time shift between consecutive blocks. Accordingly, kN parameters are

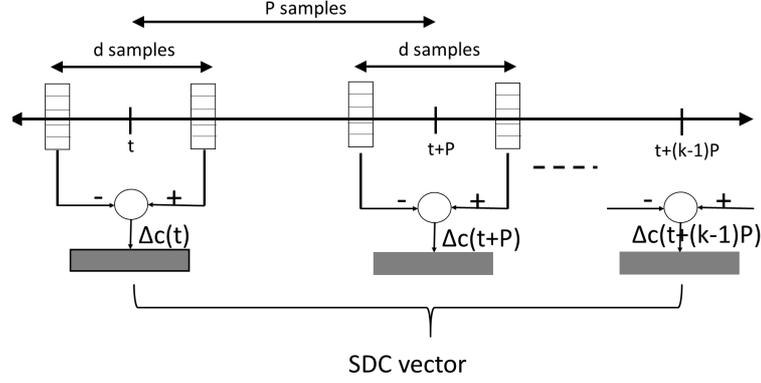


Fig. 3: Computation of SDC coefficients using MFCC and delta MFCC features.

used for each SDC feature vector, as compared with $2N$ for conventional cepstral and delta-cepstral feature vectors. The SDC is calculated as follows:

$$\Delta c(t + iP) = c(t + iP + d) - c(t + iP - d) \quad (2)$$

The final vector at time t is given by the concatenation of all $\Delta c(t + iP)$ for all $0 \leq i < k - 1$, where $c(t)$ is the original feature value at time t . In the current study, the feature vectors with static MFCC features and SDC coefficients are of length 112. The concatenated MFCC and SDC features are used to extract the i-vectors that are used in classification when applying SVM and PLDA classifiers. In the case of using CNN, the MFCC and SDC features are used as input. Figure 3 illustrates the extraction of SDC features.

2.4 Evaluation Measures

In the current study, the classification rates are used as evaluation measures. The classification rate is defined as:

$$acc = \frac{1}{n} \sum_{k=1}^n \frac{\text{No. of corrects for class } k}{\text{No. of trials for class } k} \cdot 100 \quad (3)$$

where n is the number of the emotions.

2.5 The i-vector paradigm

A widely used classification approach in speaker recognition is based on GMMs with universal background models (UBM). In this approach, each speaker model is created by adapting the UBM using maximum a posteriori (MAP) adaptation. A GMM supervector is constructed by concatenating the means of the adapted

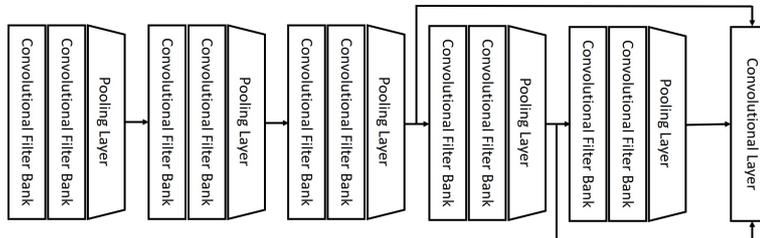


Fig. 4: The architecture of the deep feature extractor along with the classifier used during feature learning.

models. As in speaker recognition, GMM supervectors can also be used for emotion classification. The main disadvantage of GMM supervectors, however, is their high dimensionality, which imposes high computation and memory costs. In the i-vector paradigm, the limitations of high dimensional supervectors (i.e., concatenation of the means of GMMs) are overcome by modeling the variability contained in the supervectors with a small set of factors. Considering speech emotion classification, an input utterance can be modeled as:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \quad (4)$$

where \mathbf{M} is the emotion-dependent supervector, \mathbf{m} is the emotion-independent supervector, \mathbf{T} is the total variability matrix, and \mathbf{w} is the i-vector. Both the total variability matrix and emotion-independent supervector are estimated from the complete set of training data.

2.6 Classification approaches

Convolutional Neural Networks (CNN) A deep neural network is a feed-forward neural network with more than one hidden layer. The units (i.e., neurons) of each hidden layer take all outputs of the lower layer and pass them through an activation function. A convolutional neural network is a special variant of the conventional network, which introduces a special network structure. This network structure consists of alternating convolution and pooling layers.

Convolutional neural networks have been successfully applied to sentence classification [14], image classification [26], facial expression recognition [12], and in speech emotion recognition [16]. Furthermore, in [7] bottleneck features extracted from CNN are used for robust language identification.

Deep learning (DL) is behind several of the most recent breakthroughs in computer vision, speech recognition, and agents that achieved human-level performance in several games like go, poker etc. In this paper, DL for learning informative features from the signal that is then used for emotion classification is investigated. The MFCC and SDC features are calculated using overlapping windows of length 20ms. This generates multidimensional time-series that represent the data for each session. The same train/test split is used in the following

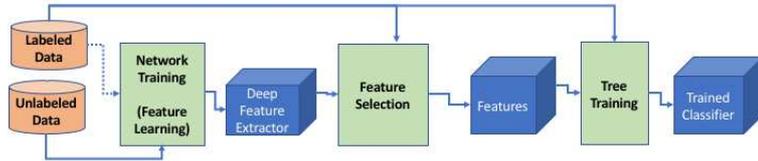


Fig. 5: The proposed training process showing the three stages of training and the output of each stage.

experiments as with the SVM and PLDA classifiers. The proposed method is a simplified version of the method recently proposed in [19] for activity recognition using mobile sensors.

The proposed classifier consists of a deep convolutional neural network (DCNN) followed by extremely randomized trees instead of the standard fully connected classifier. The motivation of using extremely randomized trees lies on previous observations showing the effectiveness in the case of small number of features. The network architecture is shown in the Figure 4, and consists of a series of five blocks, each consisting of two convolutional layers (5×64) followed by a max-pooling layer ($width = 2$). Outputs from the last three blocks are then combined and flattened to represent the learned features.

Training of the classifier proceeds in three stages as shown in the Figure 5: Network training, feature selection, and tree training. During network training, the deep convolutional neural network is trained with predefined windows of 21 feature MFCC/SDC blocks. Network training consists of two sub-stages: Firstly, the network is concatenated with its inverse to form an auto-encoder that is trained in unsupervised mode using all data in the training set and without the labels. Secondly, three fully connected layers are attached to the output of the network, and the whole combined architecture is trained as a classifier using the labeled training set. These fully connected layers are then removed, and the output of the neural network (i.e., deep feature extractor) represents the learned features.

The second training stage (i.e., feature selection) involves selecting few of the outputs from the deep feature extractor to be used in the final classification. Each feature (i.e., neuronal output i) is assigned a total *quality* ($Q(i)$) according to Equation 5, where $\bar{I}_j(i)$ is z-score normalized feature *importance* ($I_j(i)$) according to a base feature selection method.

$$Q(i) = \sum_{j=0}^{n_f} w_j \bar{I}_j(i), \quad (5)$$

In the current study, three base selectors are utilized: randomized logistic regression [6], linear SVMs with L_1 penalty, and extremely randomized trees. Random linear regression (RLR) estimates feature importance by randomly selecting subsets of training samples and fitting them using a L_1 sparsity inducing penalty that is scaled for a random set of coefficients. The features that ap-

pear repeatedly in such selections (i.e., with high coefficients) are assumed to be more *important*, and are given higher scores. The second base extractor uses a linear SVM with an L_1 penalty to fit the data and then select the features that have nonzero coefficients, or coefficients under a given threshold, from the fitted model. The third feature selector employs extremely randomized trees. During fitting decision trees, features that appear at lower depths are generally more important. By fitting several such trees, feature importance can be estimated as the average depth of each feature in the trees. Feature selection uses n -fold cross validation to select an appropriate number of neurons to keep in the final (fast) feature extractor (Figure 5). For the sake of this work, the features (outputs) that have quality (Q_i) above the median value of qualities are kept.

Given the selected features from the previous step, an extremely randomized tree classifier is then trained using the labeled data set (i.e., tree training stage).

Note that the approach described above allows to generate a classification decision for each 21 MFCC/SDC blocks. To generate a single emotion prediction for each test sample, the outputs of the classifier need to be combined. One possibility is to use a recurrent neural network, an LSTM, or HMM to do this aggregation. Nevertheless, in this work the simplest voting aggregator, in which the label of the test file is the mode of the labels of all its data, is used.

Support Vector Machine (SVM) A support vector machine is a discriminative classifier, which is widely used in regression and classification. Given a set of labeled training samples, the algorithm finds the optimal hyperplane that categorizes new samples. SVM is among the most popular machine learning methods. The advantages of SVM include the support of high-dimensionality, memory efficiency, and versatility. However, when the number of features exceeds the number of samples, the SVM performs poorly. Another disadvantage is that SVM is not probabilistic because it works by categorizing objects based on the optimal hyperplane.

Probabilistic Linear Discriminant Analysis (PLDA) PLDA is a popular technique for dimension reduction using the Fisher criterion. Using PLDA, new axes are found, which maximize the discrimination between the different classes. PLDA was originally applied to face recognition [25], and is applied successfully to specify a generative model of the i-vector representation. PLDA was also used in speaker recognition. Adapting to emotion recognition, for the i -th emotion, the i-vector $\mathbf{w}_{i,j}$ representing the j -th recording can be formulated as:

$$\mathbf{w}_{i,j} = \beta + \mathbf{S}\mathbf{x}_i + \mathbf{e}_{i,j} \quad (6)$$

where β is a global offset (i.e., mean of training vectors), \mathbf{S} represents the between-emotion variability, and the latent variable \mathbf{x} is assumed to have a standard normal distribution, and to represent a particular emotion and channel. The residual term $\mathbf{e}_{i,j}$ represents the within-emotion variability, and it is assumed to have a normal distribution with zero mean and covariance Σ .

Table 1: Average classification rates using clean models and reverberant test data (IEMOCAP).

| Reverberation time [sec] | Classification Method | | |
|--------------------------|-----------------------|------|------|
| | DCNN | SVM | PLDA |
| Clean | 71.1 | 66.1 | 62.7 |
| 0.30 | 64.3 | 48.9 | 52.5 |
| 0.47 | 64.1 | 40.7 | 50.7 |
| 0.60 | 61.1 | 35.4 | 48.6 |
| 0.78 | 64.2 | 32.5 | 46.5 |
| 1.30 | 64.3 | 35.4 | 47.2 |

After the training and test i-vectors are computed, PLDA is used to decide whether two i-vectors belong to the same class. For this task, a test i-vector and an emotion i-vector are required. The emotion i-vectors are computed as the average of the training i-vectors, which belong to a specific emotion. A classification trial requires the emotion i-vectors, the test i-vector, and the PLDA model $\{\beta, \mathbf{S}, \Sigma\}$ parameters. A closed form for PLDA scoring is presented in [27].

3 Results

This section presents the results achieved for far-field speech emotion recognition using spontaneous emotional speech data. The results presented include classification rates when using clean and reverberant data, along with DCNN, SVM, and PLDA classifiers. Furthermore, the results when using the proposed multi-style training, which addresses far-field speech emotion recognition, are demonstrated. The i-vectors dimension was set to 100, and 128 Gaussian components were used in UBM training.

Using MFCC features along with SDC features and a DCNN classifier for the clean case of using IEMOCAP data, an average classification rate of 71.1% was obtained. This result is very promising and superior to the results obtained in similar studies [11, 36]. The result also shows that SDC features can also be successfully used in speech emotion recognition. Based on this observation, in the following experiments, MFCC features concatenated with SDC coefficients are being used.

Table 1 shows the results when using clean training data and reverberant test data. As shown, in the case of SVM and PLDA classifiers and using clean models, the classification rates are decreased very significantly. On the other hand, when using DCNN, the effect of reverberation on classification rates is less significant. The results also show that PLDA classifiers outperforms SVM in the case of reverberant speech emotion recognition. On the other hand, using clean test data, SVM shows superior performance compared to PLDA. The highest rates, however, are being obtained when using DCNN.

Table 2: Average classification rates based on multi-style training and using reverberant test data (IEMOCAP).

| Reverberation time [sec] | Classification Method | | |
|--------------------------|-----------------------|------|------|
| | DCNN | SVM | PLDA |
| 0.30 | 68.3 | 66.4 | 60.6 |
| 0.47 | 68.4 | 66.4 | 61.3 |
| 0.60 | 70.4 | 66.1 | 61.3 |
| 0.78 | 71.0 | 66.1 | 63.0 |
| 1.30 | 71.1 | 65.4 | 60.6 |

Table 2 shows the classification rates when multi-style training along with reverberant test data were used in the case of IEMOCAP spontaneous emotional speech. As shown, multi-style training significantly improves the classification rates for all reverberation times. The achieved rates are almost the same as those obtained using clean data and in some cases higher rates are being obtained compared to the clean case (i.e., 0.30 sec and 0.47 sec reverberation times using SVM). A possible reason for this is the larger amount of training data used in creating the universal, reverberation-independent models. The results show the effectiveness of multi-style training in far-field speech emotion recognition, and demonstrate the ability of i-vectors to capture the variability of data with different reverberation times in a universal, reverberation-independent model set. As also shown in table 2, the DCNN classifier has superior performance compared to SVM and PLDA in the case of speech emotion recognition based on a multi-style training approach. Specifically, the classification rates obtained are closely comparable to those obtained when using clean data.

These results support the statement, that multi-style training is an effective approach to deal with the decreased performance of a speech emotion recognition system operating in far-field mode. The results also justify the previous claims, that multi-style training can successfully be applied in the full extraction of i-vectors, and that i-vectors can capture the variability of data with a large number of reverberation times and additive noise.

4 Conclusion

The current study focused on far-field speech emotion recognition using the state-of-the-art IEMOCAP spontaneous emotional database based on DCNN. Using real impulse responses convoluted with clean data and real additive noise superimposed on the convoluted data, a realistic environment for far-field speech emotion recognition was produced. In the case of using clean IEMOCAP data and DCNN, a 71.1% classification rate was obtained. This result is very promising and superior to other studies using the same databases. On the other hand, the rates were decreased in a reverberant environment when using SVM and PLDA classifiers in the i-vector paradigm. When using DCNN, the classifica-

tion rates are comparable to those obtained when using clean data. To address the problem of reverberation and additive noise, a method based on multi-style training was proposed. This resulted in the classification rates being significantly improved. The results obtained are very promising and demonstrate the effectiveness of using DCNN in far-field speech emotion recognition. Furthermore, the results show that multi-style training can be successfully applied in far-field speech emotion recognition. The effectiveness of using SDC coefficients in speech emotion recognition was also demonstrated.

References

1. Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Deng, L., Penn, G., Yu, D.: Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 1533–1545 (2014)
2. Bielefeld, B.: Language identification using shifted delta cepstrum. In *Fourteenth Annual Speech Research Symposium* (1994)
3. Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., Narayanan, S.: IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation* pp. 335–359 (2008)
4. Busso, C., Bulut, M., Narayanan, S.: Toward Effective Automatic Recognition Systems of Emotion in Speech. In: Gratch, J., Marsella, S. (eds.) *Social emotions in nature and artifact: emotions in human and human-computer interaction*, pp. 110–127. Oxford University Press, New York, NY, USA (November 2013)
5. Cristianini, N., S.-Taylor, J.: *Support vector machines*. Cambridge University Press, Cambridge (2000)
6. Friedman, J., Hastie, T., et al., R.T.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics* 28(2), 337–407 (2000)
7. Ganapathy, S., Han, K., Thomas, S., Omar, M., Segbroeck, M.V., Narayanan, S.S.: Robust Language Identification Using Convolutional Neural Network Features. in *Proc. of Interspeech* (2014)
8. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees . *Machine Learning* 63, Issue 1, 3–42 (2006)
9. Ha, H.K., Kim, N.K., Seong, W.K., Kim, H.K.: Noise-Robust Speech Emotion Recognition Using Denoising Autoencoder. *Audio Engineering Society Convention* 140 (2016), <http://www.aes.org/e-lib/browse.cfm?elib=18164>
10. Huang, C., Chen, G., Yu, H., Bao, Y., Zhao, L.: Speech Emotion Recognition under White Noise. *Archives of Acoustics* 38, 457–463 (2013)
11. Huang, C.W., Narayanan, S.: Attention Assisted Discover of Sub-Utterance in Speech Emotion Recognition. in *Proc. of Interspeech* pp. 1387–1391 (2016)
12. Huynh, X.P., Tran, T.D., Kim, Y.G.: Convolutional Neural Network Models for Facial Expression Recognition Using BU-3DFE Database . In: Kim, K., Joukov, N. (eds.) *Information Science and Applications (ICISA) 2016. Lecture Notes in Electrical Engineering*, vol. 376, pp. 441–450. Springer (2013)
13. Kanagasundaram, A., Dean, D., Sridharan, S.: Improving PLDA Speaker Verification With Limited Development Data. in *Proc. of ICASSP* pp. 1684–1688 (2014)
14. Kim, Y.: Convolutional Neural Networks for Sentence Classification. in *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 1746–1751 (2014)

15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems* 25. pp. 1097–1105. Curran Associates, Inc. (2012)
16. Lim, W., Jang, D., Lee, T.: Speech Emotion Recognition Using Convolutional and Recurrent Neural Networks. in *Proc. of Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (2016)
17. Liu, R.X.Y.: Using i-vector space model for emotion recognition. in *Proc. of Interspeech* pp. 2227–2230 (2012)
18. Metallinou, A., Lee, S., Narayanan, S.: Decision Level Combination of Multiple Modalities for Recognition and Analysis of Emotional Expression. in *Proc. of ICASSP* pp. 2462–2465 (2010)
19. Mohammad, Y., Matsumoto, K., Hoashi, K.: Deep feature learning and selection for activity recognition. In: *Proc. of the 33rd ACM/SIGAPP Symposium On Applied Computing*. pp. 926–935. ACM SAC (2018)
20. Nakamura, S., Hiyane, K., Asano, F., Endo, T.: Sound Scene Data Collection in Real Acoustical Environments. *J. Acoust. Soc. Japan (E)* 20, No.3 (1999)
21. Nicholson, J., Takahashi, K., Nakatsu, R.: Emotion Recognition in Speech Using Neural Networks. *Neural Computing & Applications* 9, Issue 4, 290–296 (2000)
22. Pan, Y., Shen, P., Shen, L.: Speech Emotion Recognition Using Support Vector Machine. *International Journal on Smart Home* 6 (2), 101–108 (2012)
23. Pohjalainen, J., Ringeval, F., Zhang, Z., Schuller, B.: Spectral and Cepstral Audio Noise Reduction Techniques in Speech Emotion Recognition. in *Proc. of ACM* (2016)
24. Prabhavalkar, R., Alvarez, R., Parada, C., Nakkiran, P., Sainath, T.: Automatic Gain Control and Multi-style Training for Robust Small-Footprint Keyword Spotting with Deep Neural Networks. in *Proc. of ICASSP* pp. 4704–4708 (2015)
25. Prince, S., Elder, J.: Probabilistic Linear Discriminant Analysis for Inferences About Identity. In *Proc. of International Conference on Computer Vision* pp. 1–8 (2007)
26. Rawat, W., Wang, Z.: Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review . *Neural Communication* 29, 23522449 (2017)
27. Romero, D.G., Wilson, C.E.: Analysis of i-vector Length Normalization in Speaker Recognition Systems. in *Proc. of INTERSPEECH* pp. 249–252 (2011)
28. Sahidullah, M., Saha, G.: Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition. *Speech Communication* 54 (4), 543565 (2012)
29. Schuller, B., Arsic, D., Wallhoff, F., Rigoll, G.: Emotion Recognition in the Noise Applying Large Acoustic Feature Sets. in *Proc. of 3rd International Conference on Speech Prosody* pp. 276–279 (2006)
30. Schuller, B., Rigoll, G., Lang, M.: Hidden Markov Model-based Speech Emotion Recognition. in *Proc. of the IEEE ICASSP I*, 401–404 (2003)
31. Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., Schuller, B.: Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks. in *Proc. of ICASSP* pp. 5688–5691 (2011)
32. Suzuki, Y., Asano, F., Kim, H., Sone, T.: An Optimum Computer-generated Pulse Signal Suitable for the Measurement of Very Long Impulse Responses. *J. Acoust. Soc. Am.* Vol. 97, No. 2, 1119–1123 (1995)
33. T.-Carrasquillo, P., Singer, E., Kohler, M.A., Greene, R.J., Reynolds, D.A., Jr., J.D.: Approaches to language identification using gaussian mixture models and

- shifted delta cepstral features. in Proc. of ICSLP2002-INTERSPEECH2002 pp. 16–20 (2002)
34. Tang, H., Chu, S., Johnson, M.H.: Emotion Recognition From Speech Via Boosted Gaussian Mixture Models. in Proc. of ICME pp. 294–297 (2009)
 35. Tawari, A., Trivedi, M.: Speech Emotion Analysis in Noisy Real-World Environment. in Proc. of International Conference on Pattern Recognition pp. 4605–4608 (2010)
 36. Tzinis, E., Potamianos, A.: Segment-Based Emotion Recognition Using Recurrent Neural Networks. in Proc. of ACII pp. 190–195 (2017)
 37. Xia, R., Liu, Y.: Using i-vector space model for emotion recognition. in Proc. of INTERSPEECH pp. 2227–2230 (2012)
 38. Zhang, T., Wu, J.: Speech emotion recognition with i-vector feature and RNN model. 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP) (2015)