Speech emotion recognition using spontaneous children's corpus

Panikos Heracleous[1], Yasser Mohammad[2], Keiji Yasuda[1,3], Akio Yoneyama[1]

[1]KDDI Research, Inc., Japan
2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502, Japan
{*pa-heracleous,yoneyama*}*@kddi-research.jp*
[2]Artificial Intelligence Research Center, AIST, Japan
yasserm@aun.edu.eg
[3]Nara Institute of Science and Technology, Japan
ke-yasuda@dsc.naist.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. Previously, several studies reported speech emotion recognition using acted emotional corpus. For real world applications, however, spontaneous corpora should be used in recognizing human emotions from speech. This study focuses on speech emotion recognition using the FAU Aibo spontaneous children's corpus. A method based on the integration of feed-forward deep neural networks (DNN) and the i-vector paradigm is proposed, and another method based on deep convolutional neural networks (DCNN) for feature extraction and extremely randomized trees as classifier is presented. For the classification of five emotions using balanced data, the proposed methods showed unweighted average recalls (UAR) of 61.1% and 59.2%, respectively. These results are very promising showing the effectiveness of the proposed methods in speech emotion recognition. The two proposed methods based on deep learning (DL) were compared to a support vector machines (SVM) based method and they demonstrated superior performance.

Keywords: Speech emotion recognition, spontaneous corpus, deep neural networks, feature extraction, extremely randomized trees

1 Introduction

Emotion recognition plays an important role in human-machine communication [1]. 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 to play in call centers in detecting a 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 is experimentally investigated.

Previous studies reported automatic speech emotion recognition using Gaussian mixture models (GMMs) [2,3], hidden Markov models (HMM) [4], support vector machines (SVM) [5], neural networks (NN) [6], and DNN [7,8]. In [9], a study based on concatenated i-vectors is reported. Audiovisual emotion recognition is presented in [10].

Previously, i-vectors were used in speech emotion recognition [11]. However, only a very few studies reported speech emotion recognition using i-vectors integrated with DNN [12]. Furthermore, to our knowledge the integration of i-vectors and DL for speech emotion recognition when limited data are available has not been investigated exhaustively so far and, therefore, the research area still remains open. Additionally, in the current study the FAU Aibo [13] state-of-the-art spontaneous children's emotional corpus is used for the classification of five emotions based on DNN and i-vectors. Another method is proposed that uses DCNN [14,15] to extract informative features, which are then used by extremely randomized trees [16] for emotion recognition. The extremely randomized trees classifier is similar to the random forest classifier [17], but with randomized tree splitting. The motivation for using extremely randomized trees lies in previous observations showing their effectiveness in the case of a small number of features, and also because of the computational efficiency.

The proposed methods based on DL are compared with a baseline classification approach. In the baseline method, i-vectors and SVM are being used. To further increase temporal information in the feature vectors, in the current study, shifted delta cepstral (SDC) coefficients [18,19] were also used along with the well-known mel-frequency cepstral coefficients (MFCC) [20].

2 Methods

2.1 Data

The FAU Aibo corpus consists of 9 hours of German speech of 51 children between the ages of 10 and 13 interacting with Sony's pet robot Aibo. Spontaneous, emotionally colored children's speech was recorded using a close-talking microphone. The data was annotated in relation to 11 emotion categories by five human labelers on a word level. In the current study, the FAU Aibo data are used for classification of the emotional states of *angry*, *emphatic*, *joyful*, *neutral*, and *rest*. To use balanced training and test data, 590 training utterances and 299 test utterances randomly selected from each emotion were used.

2.2 Feature selection

MFCC features are used in the experiments. MFCCs are very commonly used features in speech recognition, speaker recognition, emotion recognition, and language identification. Specifically, in the current study, 12 MFCCs plus energy are extracted every 10 ms using a window length of 20 ms.

SDC coefficients have been successfully used in language recognition. In the current study, the use of SDC features in speech emotion recognition is also

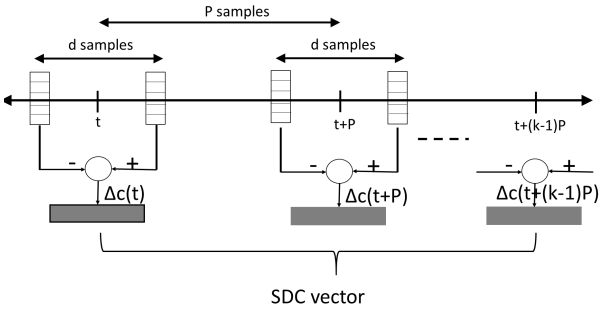

Fig. 1: Computation of SDC coefficients using MFCC and delta MFCC features.

experimentally investigated. The motivation for using SDC is to increase the temporal information in the feature vectors, which consist of frame-level features with limited temporal information. The SDC features are obtained by concatenating delta cepstral features across multiple frames. 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 used for each SDC feature vector, as compared with 2N for conventional cepstral and delta-cepstral feature vectors. The SDC is calculated as follows:

$$\Delta c(t + iP) = c(t + iP + d) - c(t + iP - d) \tag{1}$$

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 as input when using the DCNN with extremely randomized trees and conventional CNN classifiers. In the case of using DNN and SVM, the MFCC and SDC features are used to construct i-vectors used in classification. Figure 1 illustrates the extraction of SDC features.

2.3 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 models. As in speaker recognition, GMM supervectors can also be used for emotion classification.

The main disadvantage of GMM supervectors, however, is the high dimensionality, which imposes high computational 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:

$$\mathbf{M} = \mathbf{m} + \mathbf{Tw} \tag{2}$$

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.

2.4 Classification approaches

Deep Neural Networks DL [21] is behind several of the most recent breakthroughs in computer vision, speech recognition, and agents that achieved human-level performance in several games such as go, poker etc. A DNN 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. In the current study, three hidden layers with 64 units and the *ReLu* activation function are used. On top, a *Softmax* layer with five classes was added. The number of batches was set to 512, and 500 epochs were used.

Convolutional Neural Networks A CNN is a special variant of the conventional network, which introduces a special network structure consisting of alternating convolution and pooling layers. CNN have been successfully applied to sentence classification [22], image classification [23], facial expression recognition [24], and in speech emotion recognition [25]. Furthermore, in [26] bottleneck features extracted from CNN are used for robust language identification.

In this paper, DCNN 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 with a length of 20 ms. This generates a multidimensional time-series that represent the data for each session. The proposed method is a simplified version of the method recently proposed in [27] for activity recognition using mobile sensors.

The proposed classifier consists of a DCNN followed by extremely randomized trees instead of the standard fully connected classifier. The motivation for using extremely randomized trees lies in previous observations showing their effectiveness in the case of a small number of features. The network architecture is shown in Figure 2, and consists of a series of five blocks, each of which consists of two convolutional layers (64 5×5) followed by a max-pooling layer (2×2). 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 3: Network training, feature selection, and tree training. During network training, the DCNN is trained with predefined windows of 21 feature MFCC/SDC blocks (21×112 features). Network training consists of two

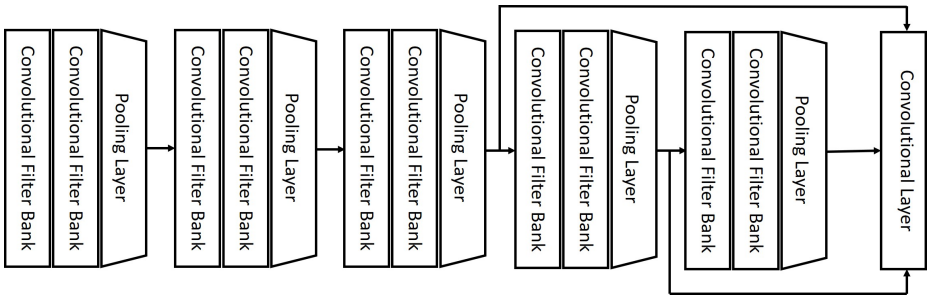

Fig. 2: The architecture of the deep feature extractor along with the classifier used during feature learning.

sub-stages: First, 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 (i.e., pre-training stage). Second, 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. Every hidden layer is an optimized classifier, and an optimized classifier is a useful feature extractor because the output is discriminative.

The second training stage (i.e., feature selection) involves selecting a 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 3, where $\bar{I}_j(i)$ is z-score normalized feature *importance* ($I_j(i)$) according to a base feature selection method.

$$Q(i) = \sum_{j=0}^{n_f} w_j \bar{I}_j(i), \tag{3}$$

In the current study, three base selectors are utilized: randomized logistic regression [28], linear SVMs with $L1$ penalty, and extremely randomized trees. Random linear regression (RLR) estimates feature importance by randomly selecting subsets of training samples and fitting them using a L1 sparsity inducing penalty that is scaled for a random set of coefficients. The features that appear 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 $L1$ 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 of 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 retain in the final

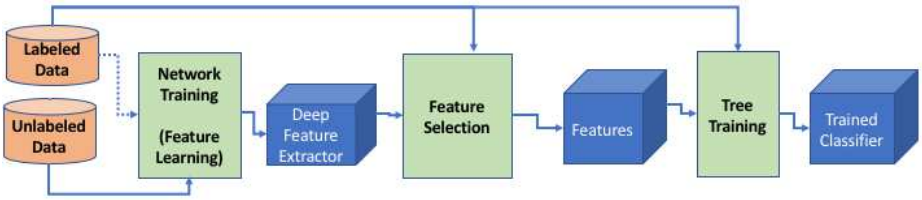

Fig. 3: The proposed training process showing the three stages of training and the output of each stage.

Table 1: Equal error rates (EER) for individual emotions when using three different classifiers.

Classifier	Angry	Emphatic	Joyful	Neutral	Rest	Average
DNN	20.1	19.8	16.4	21.1	29.8	**21.4**
DCNN + randomized trees	24.1	24.7	23.7	30.4	29.4	26.5
SVM	23.7	27.3	20.4	30.4	41.5	28.7

(fast) feature extractor (Figure 3). For this study, the features (outputs) whose quality (Q_i) exceeds the median value of qualities are retained

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 a classification decision to be generated for each of the 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 (RNN), an LSTM, or HMM to perform this aggregation. Nevertheless, in this study, the simplest voting aggregator, in which the label of the test file is the mode of the labels of all its data, is used.

3 Results

In the current study, the equal error rate (EER) and the UAR are used as evaluation measures. The UAR is defined as the mean value of the recall for each class. In addition, in the current study, the detection error tradeoff (DET) graphs are also shown.

Table 1 shows the EERs when using the three classifiers. As shown, by using DNN along with i-vectors, the lowest EER is obtained. Specifically, when using DNN, the EER was 21.4%. The second lowest EER was obtained using DCNN with extremely randomized tress. In this case, a 26.5% EER was obtained. Using SVM, the EER was 28.7%. The results also show, that *joyful*, *emphatic*, and *angry* emotions show the lowest EERs. A possible reason may be the higher

Table 2: Confusion matrix of five emotions recognition when using DNN with i-vectors.

	Angry	Emphatic	Joyful	Neutral	Rest
Angry	**63.5**	14.4	6.7	5.0	10.4
Emphatic	15.1	**63.9**	0.3	14.3	6.4
Joyful	3.7	2.3	**68.9**	4.4	20.7
Neutral	3.3	14.4	6.4	**60.2**	15.7
Rest	12.3	9.0	17.1	12.4	**49.2**

Table 3: Confusion matrix of five emotions recognition when using DCNN and extremely randomized trees.

	Angry	Emphatic	Joyful	Neutral	Rest
Angry	**65.2**	1.7	2.7	0.3	30.1
Emphatic	13.7	**61.2**	2.0	0	23.1
Joyful	4.2	2.2	**61.3**	0	32.3
Neutral	8.7	9.4	1.0	**43.8**	37.1
Rest	16.8	10.5	8.7	5.3	**58.7**

Table 4: Confusion matrix of five emotions recognition when using conventional CNN.

	Angry	Emphatic	Joyful	Neutral	Rest
Angry	**51.5**	15.1	11.4	10.0	12.0
Emphatic	10.7	**53.8**	14.7	12.7	8.1
Joyful	13.4	12.0	**51.8**	12.0	10.8
Neutral	11.7	13.4	12.4	**52.5**	10.0
Rest	13.7	8.0	16.4	9.4	**52.5**

Table 5: Confusion matrix of five emotions recognition when using SVM.

	Angry	Emphatic	Joyful	Neutral	Rest
Angry	**55.2**	15.7	6.0	7.0	16.1
Emphatic	16.7	**44.5**	3.3	17.1	18.4
Joyful	3.3	2.7	**62.2**	4.7	27.1
Neutral	7.7	12.4	13.6	**35.5**	30.8
Rest	11.4	9.4	18.7	14.0	**46.5**

emotional information included in these three emotions. On the other hand, the highest EER were obtained in the case of *neutral* and *rest* emotions (i.e., less emotional states).

The UAR when using DNN with i-vectors was 61.1%. This is a very promising result and superior to other similar studies [29–31] that used different classifiers and features with unbalanced data. The result also show that DNN and i-vectors can be effectively integrated in speech emotion recognition even in the case of limited training data. The second highest UAR was obtained in the case of DCNN with extremely randomized trees. In this case, a 59.2% UAR was achieved. When a fully-connected layer was used on top of the convolutional layers (i.e., conventional CNN classifier) the UAR was 52.4%. This rate was lower compared to the extremely randomized trees classifier with deep feature extractor. Finally, when using SVM and i-vectors, a 48.8% UAR was achieved. The results show that when using the two proposed methods based on DL, higher UARs are achieved compared to the baseline approach.

Tables 2, 3, 4, and 5 show the confusion matrices. As shown, in the case of DNN, the classification rates are comparable (with the exception of *rest*). The *joyful*, *emphatic*, *angry* classes are recognized with the highest rates, and *rest* is recognized with the lowest rate. In the case of using DCNN with extremely randomized trees, the classes *angry* and *joyful* show the highest rates. When using the conventional CNN, similar rates were obtained for all emotions. In the case of SVM, *joyful* and *angry* are recognized with the highest accuracy. It can be, therefore, concluded that the emotions *angry* and *joyful* are recognized with the highest rates in most cases.

Figures 4, 5, and 6 show the DET curves of the five individual emotions recognition. As shown, in all cases, superior performance was achieved for the emotion *joyful*.

Figure 7 shows the overall DET curves for the three classifiers. The figure clearly demonstrates that by using the two proposed methods based on DL, the highest performance is achieved. More specifically, the highest performance is obtained when using DNN and i-vectors. Note that above 30% FPR, SVM shows superior performance compared to DCNN with extremely randomized trees. The overall EER, however, is lower in the case of DCNN with extremely randomized trees compared to SVM.

4 Conclusion

The current paper focused on speech emotion recognition based on deep learning and using the state-of-the-art FAU Aibo emotion corpus of children's speech. The proposed method based on DNN and i-vectors achieved a 61.1% UAR. This result is very promising and superior to the results obtained using the same data. The results also show that i-vectors and DNN can be efficiently used in speech emotion recognition, even in the case of very limited training data. The UAR when using DCNN with extremely randomized trees was 59.2%. The two proposed methods were compared to a baseline SVM based classification scheme, and they showed superior performance. Currently, speech emotion recognition using the proposed methods and the FAU Aibo data in noisy and reverberant environments is being investigated.

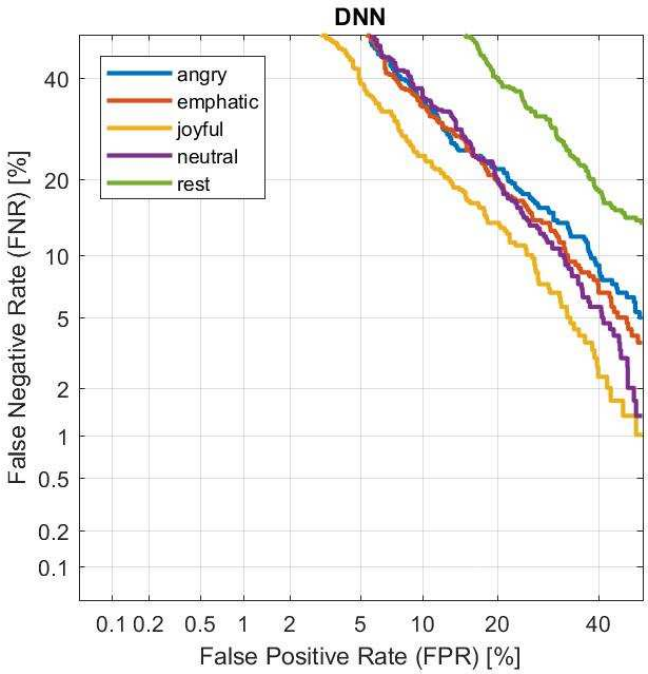

Fig. 4: DET curves of speech emotion recognition using DNN.

References

1. C. Busso, M. Bulut, and S. Narayanan, "Toward Effective Automatic Recognition Systems of Emotion in Speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.
2. H. Tang, S. Chu, and M. H. Johnson, "Emotion Recognition From Speech Via Boosted Gaussian Mixture Models," *in Proc. of ICME*, pp. 294–297, 2009.
3. S. Xu, Y. Liu, and X. Liu, "Speaker Recognition and Speech Emotion Recognition Based on GMM," *3rd International Conference on Electric and Electronics (EEIC 2013)*, pp. 434–436, 2013.
4. B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov Model-based Speech Emotion Recognition," *in Proc. of the IEEE ICASSP*, vol. I, pp. 401–404, 2003.
5. Y. Pan, P. Shen, and L. Shen, "Speech Emotion Recognition Using Support Vector Machine," *International Journal on Smart Home*, vol. 6 (2), pp. 101–108, 2012.
6. J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion Recognition in Speech Using Neural Networks," *Neural Computing & Applications*, vol. 9, Issue 4, pp. 290–296, 2000.
7. A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke1, G. Meier, and B. Schuller, "Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks," *in Proc. of ICASSP*, pp. 5688–5691, 2011.

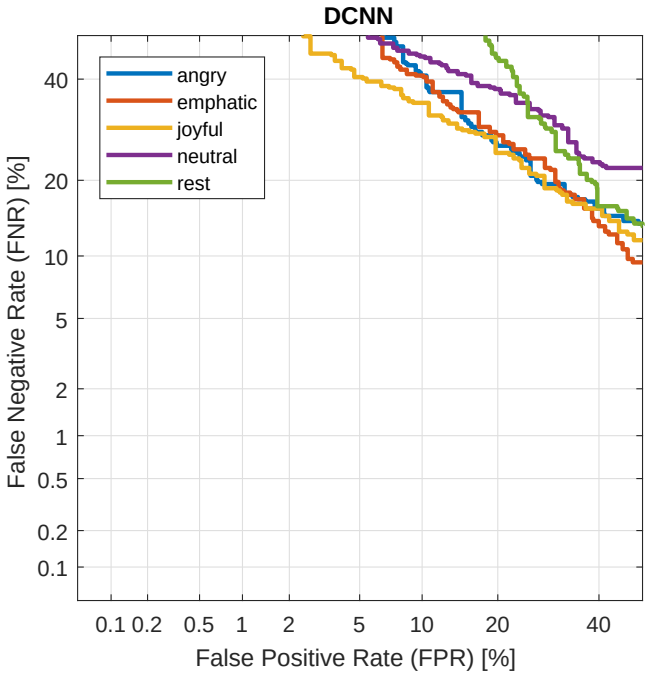

Fig. 5: DET curves of speech emotion recognition using DCNN and extremely randomized trees.

8. K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," *in Proc. of Interspeech*, pp. 2023–2027, 2014.
9. R. Xia and Y. Liu, "Using i-vector space model for emotion recognition," *in Proc. of INTERSPEECH*, pp. 2227–2230, 2012.
10. A. Metallinou, S. Lee, and S. Narayanan, "Decision Level Combination of Multiple Modalities for Recognition and Analysis of Emotional Expression," *in Proc. of ICASSP*, pp. 2462–2465, 2010.
11. R. X. Y. Liu, "Using i-vector space model for emotion recognition," *in Proc. of Interspeech*, pp. 2227–2230, 2012.
12. T. Zhang and J. Wu, "Speech emotion recognition with i-vector feature and RNN model," *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, pp. 524–528, 2015.
13. S. Steidl, "Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech," *Logos Verlag, Berlin*, 2009.
14. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
15. O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on*

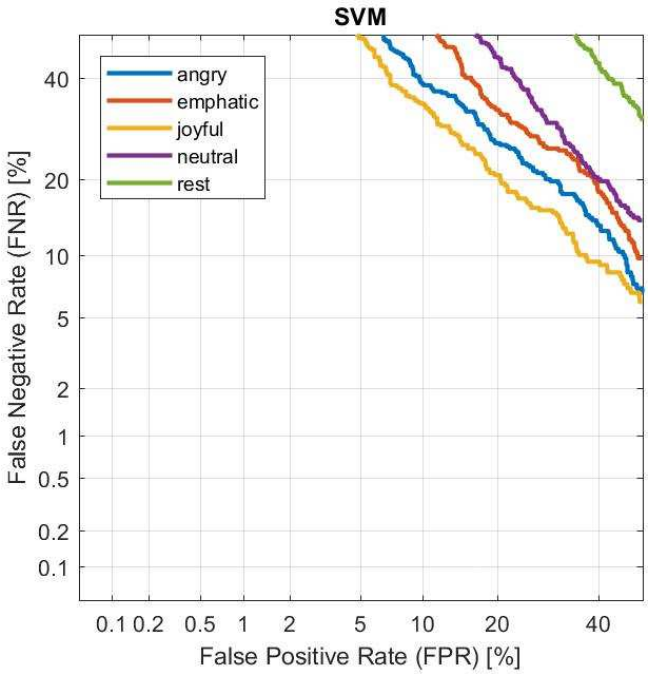

Fig. 6: DET curves of speech emotion recognition using SVM.

Audio, Speech, and Language Processing, vol. 22, pp. 1533–1545, 2014.

16. P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees ," *Machine Learning*, vol. 63, Issue 1, pp. 3–42, 2006.
17. T. K. Ho, "Random decision forests," *In Proc. of the 3rd International Conference on Document Analysis and Recognition*, pp. 278–282, 1995.
18. B. Bielefeld, "Language identification using shifted delta cepstrum," *In Fourteenth Annual Speech Research Symposium*, 1994.
19. P. T.-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, , and J. D. Jr., "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," *in Proc. of ICSLP2002-INTERSPEECH2002*, pp. 16–20, 2002.
20. M. Sahidullah and G. Saha, "Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition," *Speech Communication*, vol. 54 (4), p. 543565, 2012.
21. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups ," *IEEE Signal Processing Magazine*, vol. 29, Isuue:6, pp. 82–97, 2012.
22. Y. Kim, " Convolutional Neural Networks for Sentence Classification," *in Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014.

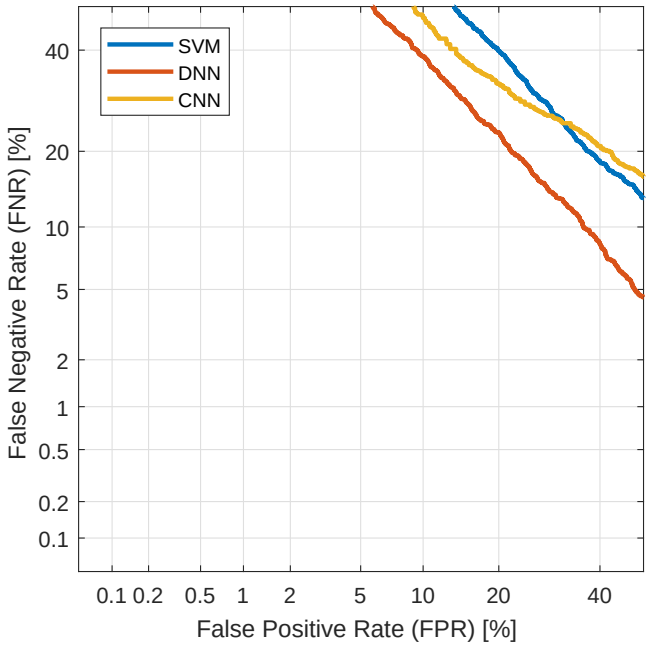

Fig. 7: DET curves of speech emotion recognition using three different classifiers.

23. W. Rawat and Z. Wang, "Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review ," *Neural Communication*, vol. 29, p. 23522449, 2017.
24. X.-P. Huynh, T.-D. Tran, and Y.-G. Kim, "Convolutional Neural Network Models for Facial Expression Recognition Using BU-3DFE Database ," in *Information Science and Applications (ICISA) 2016. Lecture Notes in Electrical Engineering*, K. Kim and N. Joukov, Eds. Springer, 2013, vol. 376, pp. 441–450.
25. W. Lim, D. Jang, and T. Lee, "Speech Emotion Recognition Using Convolutional and Recurrent Neural Networks," *in Proc. of Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016.
26. S. Ganapathy, K. Han, S. Thomas, M. Omar, M. V. Segbroeck, and S. S. Narayanan, " Robust Language Identification Using Convolutional Neural Network Features," *in Proc. of Interspeech*, 2014.
27. Y. Mohammad, K. Matsumoto, and K. Hoashi, "Deep feature learning and selection for activity recognition," in *in Proc. of the 33rd ACM/SIGAPP Symposium On Applied Computing*, ser. ACM SAC, 2018, pp. 926–935.
28. J. Friedman, T. Hastie, and R. T. et al., "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.
29. Y. Attabi, J. Alam, P. Dumouchel, P. Kenny, and D. O. Shaughnessy, "Multiple windowed spectral features for emotion recognition ," *in Proc. of ICASSP*, pp. 7527–7531, 2013.

30. H. Cao, R. Verma, and A. Nenkova, "Combining ranking and classification to improve emotion recognition in spontaneous speech ," *in Proc. of INTERSPEECH*, 2012.
31. D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks ," *in Proc. of IEEE ASRU*, pp. 216–221, 2013.