Feature Selection and Nuisance Attribute Projection for Speech Emotion Recognition

Man-Wai MAK

Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University
enmwmak@polyu.edu.hk

Abstract

In speech emotion recognition, we need to maximize the variability of emotion features across emotion states and suppress the non-emotion variabilities. This paper investigates methods to achieve both of these goals. Specifically, utterance-level emotion features are extracted and the most relevant features are selected by ranking their Fisher discriminant ratios; nuisance attribute projection (NAP) is applied to project the emotion vectors to a subspace with minimum non-emotion variability. The projected vectors are classified by support vector machines or deep neural networks. Evaluations on Emo-DB and CHEAVD suggest that (1) among the 4000+ features defined in the Interspeech speaker-state challenge, many of them have low discriminative power and can be dropped without affecting performance; (2) NAP can remove some of the non-emotion variability; and (3) it is better to use a richer set of features and let the feature selection algorithm to select the relevant ones rather than starting with a compact feature set without feature selection.[1]

Keywords: Emotion recognition, nuisance attribute projection, support vector machines, deep neural networks

[1]Citing this paper: M.W. Mak, "Feature Selection and Nuisance Attribute Projection for Speech Emotion Recognition", *Technical Report and Lecture Note Series, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University*, Dec. 2016.

1. Introduction

Human has the intrinsic ability to recognize the emotion of people. Indeed, without this ability, human interaction will become very difficult. To make human computer interaction more natural, it is important to develop software that can recognize the emotion of human. In the case where only audio signals are available, the recognition and tracking of emotion can only be achieved by analyzing the speech signals. A typical example of such scenario is spoken dialog systems for customer services.

In recent years, much progress in speech emotion recognition has been made. Much of the effort has been spent on finding useful local features (such as pitch, spectral envelope and MFCCs) from which utterance-level features are extracted through some statistical functions [1, 2]. The utterance-level features were than classified by statistical classifiers such as support vector machines (SVMs) and Gaussian mixture models [3, 4]. To improve classification performance, some authors [5, 6, 7] applied feature selection techniques, such as information gain and minimal-redundancy-maximal-relevance [7] to find the spectral and prosodic features that are mostly correlated to emotion states of speakers.

The DNN approach to emotion recognition has started to become popular very recently. Instead of using hand-craft spectral and prosodic features, deep belief networks can be used for generating emotion features from prosodic and spectral features; then, the generated features can be classified by SVMs [8]. In [9], contextual acoustic information from spectrograms were used for training a denoising autoencoder [10]; after training, emotion features were extracted from the bottleneck layer of the autoencoder [9]. To classify the frame-based bottleneck features or to exploit the dynamic structure of frame-based features, long short-term memory recurrent neural networks (LSTM-RNN) [11] have been used [9, 12, 13]. In a similar strategy, a denoising autoencoder with two sets of hidden neurons is trained to learn the hidden representation of neutral and emotional speech, respectively; then, emotion features are extracted from the emotion hidden neurons of the autoencoder [14].

In speech emotion recognition, all non-emotion variabilities embedded in the speech signals are considered as nuisance information and should be suppressed. One source of nuisance variabilities comes from the speakers, primarily because the voices of individuals are different and different persons will express the same emotion differently. While speaker variability can be normalized by factor analysis (FA) techniques [15], the method requires speaker labels in the training corpus. In this paper, we propose using

nuisance attribute projection (NAP) [16, 17, 18], which has been very successful in speaker verification, to suppress the non-emotion variability in the utterance-level feature vectors. In speaker verification, NAP finds a nuisance subspace in which most of non-speaker information varies and projects the speaker-dependent GMM-supervectors to a subspace that is orthogonal to the nuisance subspace. In emotion recognition, speaker variability becomes the nuisance and emotion variability is the variability that we want to keep. NAP has advantages over the FA approach in that it does not require speaker labels in the training corpus.

This work has investigated two utterance-level feature sets for emotion recognition. One set has over 300 features and another one has over 4,000. For the former, a majority of the features are important for emotion recognition. For the latter, some of the 4000+ may be redundant and some of them are more relevant for emotion recognition than the others. By ranking the relevance of individual features, we demonstrate that the relevance drops exponentially fast and that many of the 4000+ features have a low relevance. This observation suggests that it is important to select relevant features for emotion recognition. This paper attempts to answer the question: Shall we use a more compact feature set (300+) without feature selection or shall we select relevant features from a bigger and potentially richer feature set? Our results clearly suggest that the latter approach is better.

We applied NAP and feature selection to the utterance-level features and present the resulting feature vectors to SVM-based and DNN-based classifiers. Performance of these emotion classifiers was evaluated on two datasets: Berlin Emo-DB and CHEAVD. It was found that NAP and feature selection can improve classification performance.

2. Nuisance Attribute Projection

Nuisance attribute projection (NAP) is originally developed for removing channel effects on GMM-SVM speaker verification [16, 17, 18]. The idea is to find a subspace within the GMM-supervector space [19] in which all non-speaker variabilities occur. The method requires a training set comprising multiple speakers and multiple recording sessions per speaker.

In this work, we applied NAP to suppress the non-emotion variability in the emotion features. Unlike speaker recognition, speaker variability is one of the nuisance attributes that we want to remove. While speaker variability can be suppressed by linear discriminate analysis (LDA) and probabilistic LDA [20, 15], these methods require an emotion speech database with speaker labels. Also, To ensure sufficient rank in the covariance matrices,

these methods also require a large number of speakers in the database. On the other hand, to suppress non-emotion variability (including speaker variability), NAP only requires an emotion database with an emotion label for each utterance.

Given a training set comprising N emotion vectors $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, we aim to find a subspace defined by the column vectors in $\mathbf{V}$ such that the projected vectors

$$\tilde{\mathbf{x}}_i = \mathbf{P}\mathbf{x}_i = (\mathbf{I} - \mathbf{V}\mathbf{V}^{\mathsf{T}})\mathbf{x}_i, \quad i = 1, \ldots, N \tag{1}$$

retain most of the emotion information but with non-emotion information suppressed. For $\tilde{\mathbf{x}}_i$'s to retain most of the emotion information, the rank of $\mathbf{V}$ should be low. The subspace can be found by minimizing the objective function:

$$\mathbf{V}^* = \underset{\mathbf{V}:\mathbf{V}^{\mathsf{T}}\mathbf{V}=\mathbf{I}}{\arg\min} \sum_{ij} w_{ij} \|\mathbf{P}(\mathbf{x}_i - \mathbf{x}_j)\|^2 \tag{2}$$

where

$$w_{ij} = \begin{cases} 1 & \text{if emotion}(\mathbf{x}_i) = \text{emotion}(\mathbf{x}_j) \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Eq. 1 and Eq. 3 suggest that we pull the projected emotion vectors belong to the same emotion together and do not care about the vector pairs of different types.

To find $\mathbf{V}^*$ in Eq. 2, we define the data matrix $\mathbf{X} \equiv [\mathbf{x}_1 \ \ldots \ \mathbf{x}_N]$ and the vector difference $\mathbf{d}_{ij} \equiv (\mathbf{x}_i - \mathbf{x}_j)$. Then, the weighted projected distance in Eq. 2 can be expressed as

$$\begin{aligned} D &= \sum_{ij} w_{ij} \left(\mathbf{P}\mathbf{d}_{ij}\right)^{\mathsf{T}} \left(\mathbf{P}\mathbf{d}_{ij}\right) \\ &= \sum_{ij} w_{ij} \mathbf{d}_{ij}^{\mathsf{T}} (\mathbf{I} - \mathbf{V}\mathbf{V}^{\mathsf{T}})^{\mathsf{T}} (\mathbf{I} - \mathbf{V}\mathbf{V}^{\mathsf{T}}) \mathbf{d}_{ij} \\ &= \sum_{ij} w_{ij} \mathbf{d}_{ij}^{\mathsf{T}} \mathbf{d}_{ij} - \sum_{ij} w_{ij} \mathbf{d}_{ij}^{\mathsf{T}} \mathbf{V}\mathbf{V}^{\mathsf{T}} \mathbf{d}_{ij}, \end{aligned} \tag{4}$$

where we have used the constraint $\mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$. Dropping terms independent

of $\mathbf{V}$, we have

$$\begin{aligned} D' &= -\sum_{ij} w_{ij}(\mathbf{x}_i - \mathbf{x}_j)^\mathsf{T}\mathbf{V}\mathbf{V}^\mathsf{T}(\mathbf{x}_i - \mathbf{x}_j) \\ &= -\sum_{ij} w_{ij}\mathbf{x}_i^\mathsf{T}\mathbf{V}\mathbf{V}^\mathsf{T}\mathbf{x}_i - \sum_{ij} w_{ij}\mathbf{x}_j^\mathsf{T}\mathbf{V}\mathbf{V}^\mathsf{T}\mathbf{x}_j \\ &\quad + 2\sum_{ij} w_{ij}\mathbf{x}_i^\mathsf{T}\mathbf{V}\mathbf{V}^\mathsf{T}\mathbf{x}_j \\ &= -2\sum_{ij} w_{ij}\mathbf{x}_i^\mathsf{T}\mathbf{V}\mathbf{V}^\mathsf{T}\mathbf{x}_i + 2\sum_{ij} w_{ij}\mathbf{x}_i^\mathsf{T}\mathbf{V}\mathbf{V}^\mathsf{T}\mathbf{x}_j. \end{aligned} \tag{5}$$

Using the identity $\mathbf{a}^\mathsf{T}\mathbf{B}\mathbf{B}^\mathsf{T}\mathbf{a} = \mathrm{Tr}\{\mathbf{B}^\mathsf{T}\mathbf{a}\mathbf{a}^\mathsf{T}\mathbf{B}\}$, where Tr stands for matrix trace, Eq. 5 can be written as

$$\begin{aligned} D' &= -2\mathrm{Tr}\left\{\sum_{ij} w_{ij}\mathbf{V}^\mathsf{T}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\mathbf{V}\right\} + 2\mathrm{Tr}\left\{\sum_{ij} w_{ij}\mathbf{V}^\mathsf{T}\mathbf{x}_i\mathbf{x}_j^\mathsf{T}\mathbf{V}\right\} \\ &= -2\mathrm{Tr}\left\{\mathbf{V}^\mathsf{T}\mathbf{X}\mathrm{diag}(\mathbf{W}\mathbf{1})\mathbf{X}^\mathsf{T}\mathbf{V}\right\} + 2\mathrm{Tr}\left\{\mathbf{V}^\mathsf{T}\mathbf{X}\mathbf{W}\mathbf{X}^\mathsf{T}\mathbf{V}\right\} \\ &= 2\mathrm{Tr}\left\{\mathbf{V}^\mathsf{T}\mathbf{X}\left[\mathbf{W} - \mathrm{diag}(\mathbf{W}\mathbf{1})\right]\mathbf{X}^\mathsf{T}\mathbf{V}\right\}, \end{aligned} \tag{6}$$

where $\mathbf{W}$ is the weight matrix in Eq. 3 and $\mathrm{diag}(\mathbf{a})$ means converting $\mathbf{a}$ into a diagonal matrix, and $\mathbf{1}$ is a vector of all 1's. It can be shown that minimizing D' in Eq. 6 with the constraint $\mathbf{V}^\mathsf{T}\mathbf{V}$ is equivalent to finding the first K eigenvectors with the smallest eigenvalues of [21]:

$$\mathbf{X}\left[\mathbf{W} - \mathrm{diag}(\mathbf{W}\mathbf{1})\right]\mathbf{X}^\mathsf{T}\mathbf{V} = \mathbf{\Lambda}\mathbf{V}.$$

3. Fisher Discriminant Ratio

Fisher Discriminant Ratio [22, 23] is a feature selection method that ranks the relevance of individual features based on their degree of separability between the positive (+) and negative (−) classes:

$$\mathrm{FDR}(j) = \frac{(\mu_j^+ - \mu_j^-)^2}{\left(\sigma_j^+\right)^2 + \left(\sigma_j^-\right)^2}. \tag{7}$$

In Eq. 7, j is the feature indexes and μ_j and σ_j are the mean and standard deviation of feature j, respectively. To apply FDR for feature selection,

FDR(j)'s are sorted in descending order and the first F features in the sorted list are considered as the relevant features.

Eq. 7 define the criterion for ranking features in binary classification problems. For multi-class problems, a simple extension is to consider each class in turn and rank the feature for each class using the one-vs-rest approach. The final set of features comprise the union of the class-dependent features.

4. Experiments

4.1. Speech Data and Evaluation Protocols

The experiments were based on the Berlin Database of Emotional Speech (Emo-DB) [24] and the Chinese natural emotional audio–visual database (CHEAVD) [12].

Emo-DB contains seven categories of emotional speech spoken by ten speakers. All speakers spoke the same set of verbal content in an anechoic chamber. The database comprises 535 speech files. Because Emo-DB does not divide the data into training and test set, we performed leave-one-speaker out cross validation (LOSOCV) on the data.

CHEAVD contains 140 minutes of emotional segments extracted from Chinese movies, TV plays and talk shows. It contains 2,322 speech segments (files) spoken by 238 speakers. Each file has multiple emotional labels and one of which is the primary label. There are a total of 8 primary emotional states. In this work, we chose the 6 major categories – Angry, Disgust, Happy, Neutral, Sad, and Surprise – so that we can compare our results with those in [12]. The database divides the speech segments into three sets: training, validation and test. This work used the training and test sets.

4.2. Emotion Features

We used OpenSmile V2.1 [25] to extract emotion features from the speech files. We used the feature sets specified in Interspeech 2009 Emotion Challenge (`IS09_emotion.conf`) [1] and Interspeech 2011 Speaker State Challenge (`IS11_speaker_state.conf`) [2]. The former contains 384 features and the latter has 4,370. These feature sets comprise low-level descriptors, such as root-mean-square frame energies, MFCCs, zero-crossing rates, voice probabilities, and fundamental frequencies. The low-level descriptors are processed by statistical functionals – such as maximum, minimum, range, standard deviation, kurtosis, skewness, slope of contour, etc. – to extract

high-level descriptions of the speech signals. For both datasets, we removed the features with zero variances.

The features were normalized independently by z-norm. Then, FDR (Eq. 7) was applied to select F relevant features per class. Then, the final set of features are the union of the F FDR-selected features across all emotion classes. As a result the total number of selected features are larger than the specified cut-off F.

4.3. SVM and DNN Emotion Classifiers

We used support vector machines (SVMs) and deep neural networks (DNNs) for classification. For the SVMs, we adopted a one-vs-one approach to implementing the classifier.[2] Although the number of features in `IS11_speaker_state.conf` is larger than the feature dimension, we observe that RBF-SVMs perform slightly better than linear SVM. Therefore, we used RBF-SVM with the kernel parameters set to `1/no_of_features`.[3]

Each DNN classifier in this work comprises two hidden layers (each with 100 nodes) and a softmax output layer with either 6 or 7 output nodes (depending on the dataset used). The ReLU activation function was used for all hidden nodes. A dropout rate of 20% was applied to all layers, and L2 kernel regularization with a weight of 0.1 was applied to the second hidden layer and the output layer (just before the softmax). The Adam optimizer in the Keras library was used for training the DNNs for 20 epochs. All of the parameters in the Adam optimizer were set to their default values.

5. Results

5.1. Feature Sets and Feature Selection

Table 1 shows the performance of SVMs and DNNs using different feature sets. Evidently, IS11_Speaker_State is a better feature set. The result is reasonable because IS11_Speaker_State was defined three years after IS09_Emotion and it contains over 4,000 features, whereas IS09_Emotion only has a few hundred features.

Fig. 1 shows the histogram of FDR for the 'Angry' class against the remaining 5 classes in CHEAVD. The results suggest that while the number of features in Interspeech 2011 Speaker State Challenge is very large, many of them are not very useful and only a third of them have high FDR scores.

[2] We use the Python function `sklearn.svm.SVC()`.

[3] Python code: `svc = SVC(C=10, gamma='auto', kernel='rbf')`.

Fig. 2 further confirms this observation. The figures suggests that we may drop many features with low FDRs.

Table 2 shows the performance of SVMs and DNNs for different numbers of selected features. Because IS09_emotion only has 382 features, we focused on IS11_Speaker_State. For CHEAVD, feature selection does have significant effect on recognition performance. Using the full feature set that comprises 4,368 features degrades the performance. The accuracies in Table 2(a) are higher than the 53% achieved by LSTM-RNN in Table 7 of [12]. The superior performance of SVMs suggests that it may not be necessary to use complicated models such as LSTM-RNN for this dataset.

Fig. 3 shows the confusion matrices obtained by the best performing SVMs in Table 2. It shows that for CHEAVD, only 'Angry' and 'Neutral' are less confusable with the other emotion states. Many samples from 'Happy', 'Sad', and 'Surprise' were incorrectly classify to 'Neutral'. On the other hand, the utterances in Emo-DB are less confusable, as most of the samples were correctly classified.

5.2. Effects of NAP

Because there are only 10 speakers in Emo-DB, there is not much speaker variability in the data. Therefore, we applied NAP to CHEAVD only. CHEAVD contains the speech of 234 speakers, which have sufficient speaker variability for us to see the effect of NAP.

Table 3 shows the performance of NAP with different numbers of columns in $\mathbf{V}$, i.e., the number of dimensions to be projected out. The results show that even for 234 speakers, performance can only be improved by projecting out one nuisance dimension.

5.3. Compare with Existing Systems

Table 4 shows the performance of the proposed emotion classifiers and other existing classifiers in the literatures. Because CHEAVD was published in September 2016, there is only one paper (from the authors of CHEAVD) using this dataset. For CHEAVD, our best performing system (NAP + DNN) outperforms the one in [12] by 8.5%. For the Emo-DB, our best system is superior to the GMM and SVM classifiers in [4] but still not comparable with the performance achieved by human.

6. Conclusions

This paper proposes using nuisance attribute projection (NAP) to suppress non-emotion variability in utterance-level emotion vectors and using

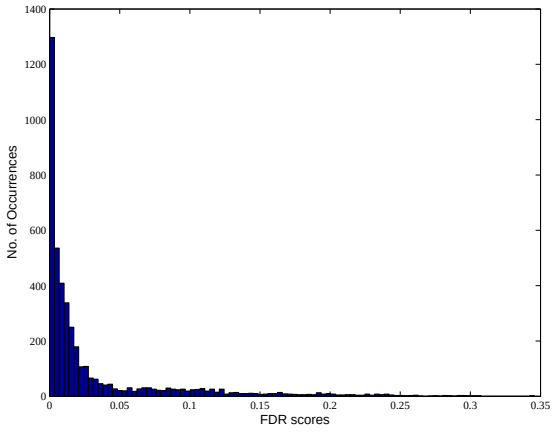


Figure 1: Distribution of FDR for the 'Angry' class against the remaining 5 classes in CHEAVD using the IS11 Speaker State feature set.

Fisher discriminant ratio (FDR) to select relevant features for SVM and DNN classifiers. Some of the observations are highlighted below:

- FDR and SVM work very well together. With FDR, the accuracy of our SVM-based emotion classifier increases by 9.3%.
- The ranked relevances of the emotion features decrease exponentially fast. As a result, many of the features can be dropped without affecting performance.
- As long as a good set of utterance-level features can be found, simple SVM classifiers can outperform the more sophisticated classifiers such as LSTM-RNN that operate on the frame or segment level.
- For small datasets, SVM classifiers have less parameters to tune and can easily beat DNN classifiers.

References

[1] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge.," in *Proc. Interspeech*, 2009, vol. 2009, pp. 312–315.

[2] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge.," in *Proc. Interspeech*, 2011, pp. 3201–3204.

[3] H. W. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, 2015.

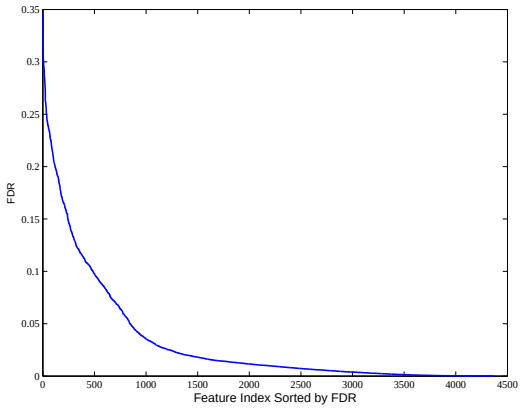


Figure 2: FDR scores ('Angry' vs the remaining 5 classes) sorted in descending order.

Features Set	SVM	DNN
IS09_Emotion (Full)	50.44%	56.19%
IS11_Speaker_State (Full)	52.21%	55.75%

(a) CHEAVD

Features Set	SVM	DNN
IS09_Emotion (Full)	75.70%	73.83%
IS11_Speaker_State (Full)	80.56%	80.19%

(b) Burlin Emo-DB

Table 1: Accuracy of SVMs and DNNs using different feature sets without feature selection and NAP.

[4] I. Luengo, E. Navas, and I. Hernáez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 490–501, 2010.

[5] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[6] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.

[7] H. C. Peng, F. H. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,"

No. of Selected Features	SVM	DNN
542	54.87%	56.64%
1330	57.08%	57.08%
1992	55.75%	57.08%
2960	55.75%	56.64%
3599	54.42%	56.64%
4368 (Full)	52.21%	55.75%

(a) CHEAVD

No. of Selected Features (F) per Class	SVM	DNN
500	81.50%	80.00%
1000	81.86%	80.37%
1500	80.37%	79.44%
2000	80.93%	79.25%
2500	80.56%	78.88%
3000	80.37%	80.75%
Full	80.56%	80.19%

(b) Burlin Emo-DB

Table 2: (a) Accuracy of SVMs and DNNs on (a) CHEAVD and (b) Emo-DB using different number of features (F) from IS11 Speaker State. F = Full means using the full feature set.

IEEE Transactions on pattern analysis and machine intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.

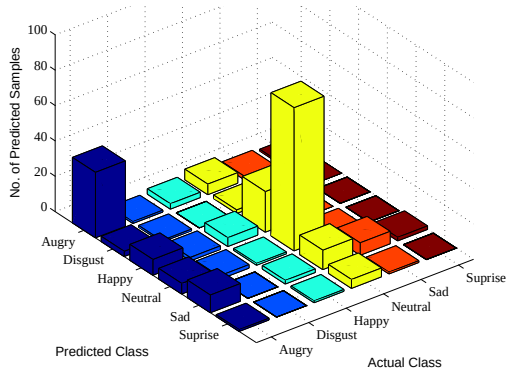
[8] Y. Kim, H. L. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3687–3691.

[9] S. Ghosh, E. Laksana, L. P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," *Proc. Interspeech*, pp. 3603–3607, 2016.

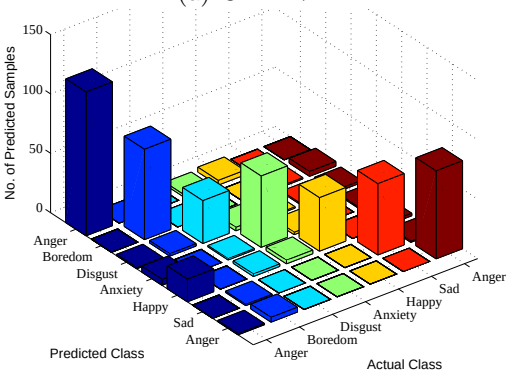
[10] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. ICML*, 2008, pp. 1096–1103.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] Y. Li, J. H. Tao, L. L. Chao, W. Bao, and Y. Z. Liu, "CHEAVD: a chinese natural emotional audio–visual database," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2016.



(a) CHEAVD



(b) Burlin Emo-DB

Figure 3: (a) Confusion matrix obtained from the SVM classifier in (a) Table 2(a) using 1000 features per class and (b) Table 2(b) using the full feature set.

[13] J. K. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. Interspeech*, 2015.

[14] R. Xia and Y. Liu, "Using denoising autoencoder for emotion recognition.," in *Proc. Interspeech*, 2013, pp. 2886–2889.

[15] T. Dang, V. Sethu, and E. Ambikairajah, "Factor analysis based speaker normalization for continuous emotion prediction," in *Proc. Interspeech*, 2016, pp. 913–917.

[16] A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel compensation for SVM speaker recognition," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Toledo, Spain, May–June 2004, pp. 57–62.

[17] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. of ICASSP'05*, 2005, pp. 629–632.

NAP Dim (Q)	IS11 Speaker State	
	SVM	DNN
0	52.21%	55.75%
1	53.10%	57.52%
2	51.33%	53.87%
3	52.21%	54.87%
4	51.77%	52.65%

Table 3: Accuracy of SVMs and DNNs using different numbers of NAP dimension (Q) on the test set of CHEAVD. $Q = 0$ means no NAP was applied.

Reference	Feature	Classification	Acc. (%)
Li *et al.* [12]	YAAFE	LSTM-RNN	53.00
This paper	IS11_Speaker_State w/ NAP	DNN	57.52

(a) CHEAVD

Reference	Feature	Classification	Acc. (%)
Burkhardt *et al.* [24]	Perception test by human		87.50
Luengo *et al.* [4]	Spectral & prosodic	GMM/SVM	78.30
This paper	IS11_Speaker_State w/ FDR	SVM	81.86

(b) Berlin Emo-DB

Table 4: Comparing performance with other methods in the literatures. Note that [3] used 6 emotion classes instead of 7.

[18] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, Toulouse, France, May 2006, vol. 1, pp. 97–100.

[19] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[20] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–8.

[21] E. Kokiopoulou, J. Chen, and Y. Saad, "Trace optimization and eigenproblems in dimension reduction methods," *Numerical Linear Algebra with Applications*, vol. 18, no. 3, pp. 565–602, 2011.

[22] P. Pavlidis, J. Weston, J. S. Cai, and W. N. Grundy, "Gene functional clas-

sification from heterogeneous data," in *Proc. of the 5th Annual International Conference on Computational biology*. ACM, 2001, pp. 249–255.

[23] M. W. Mak and S. Y. Kung, "Fusion of feature selection methods for pairwise scoring SVM," *Neurocomputing*, vol. 71, no. 16-18, pp. 3104–3113, 2008.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech.," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.

[25] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.