Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Speech emotion recognition with unsupervised feature learning*

Zheng-wei HUANG, Wen-tao XUE, Qi-rong MAO‡

(*Department of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China*)

E-mail: zhengwei.hg@gmail.com; striveyou@163.com; mao_qr@mail.ujs.edu.cn

Received Sept. 16, 2014; Revision accepted Mar. 4, 2015; Crosschecked Apr. 10, 2015

Abstract: Emotion-based features are critical for achieving high performance in a speech emotion recognition (SER) system. In general, it is difficult to develop these features due to the ambiguity of the ground-truth. In this paper, we apply several unsupervised feature learning algorithms (including K-means clustering, the sparse auto-encoder, and sparse restricted Boltzmann machines), which have promise for learning task-related features by using unlabeled data, to speech emotion recognition. We then evaluate the performance of the proposed approach and present a detailed analysis of the effect of two important factors in the model setup, the content window size and the number of hidden layer nodes. Experimental results show that larger content windows and more hidden nodes contribute to higher performance. We also show that the two-layer network cannot explicitly improve performance compared to a single-layer network.

Key words: Speech emotion recognition, Unsupervised feature learning, Neural network, Affect computing
doi:10.1631/FITEE.1400323 **Document code:** A **CLC number:** TP391.4

1 Introduction

Emotion recognition has attracted a lot of researchers in pattern recognition and machine learning. Most previous work on emotion recognition has focused on vocal (Pantic *et al.*, 2008) and facial (Gunes and Schuller, 2013) or other modalities like gestural effect (Nicolaou *et al.*, 2011) in terms of basic emotions. In this paper, we focus on speech emotion recognition (SER). SER is being applied to many areas such as driving safety, call centers, diagnostic tools for therapists, and especially in the situation where natural human-machine interaction is required (El Ayadi *et al.*, 2011). However, recognizing emotions from speech is a very challenging work, primarily due to the ambiguity of the ground-truth: different people may express emotions in different ways.

Much of the actual effort in deploying systems of SER goes into the design of an appropriate representation of speech signals. Affect-related features are critical for achieving high performance. The four most commonly used types of features in the literature are: (1) acoustic features, (2) linguistic features (words and discourse), (3) context information (e.g., subject, gender, and turn-level features representing local and global aspects of the dialogue) (El Ayadi *et al.*, 2011), and (4) hybrid features that use both acoustic and linguistic information. Although a number of speech emotion features have been proposed, there is not yet a manually designed optimal feature set. Researchers are likely to combine more and more features, which may cause a problem of high dimensionality. Moreover, some complex factors like the variation of speaker, content, and

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61272211 and 61170126) and the Six Talent Peaks Foundation of Jiangsu Province, China (No. DZXX027)

ORCID: Zheng-wei HUANG, http://orcid.org/0000-0001-7788-0526; Qi-rong MAO, http://orcid.org/0000-0002-5021-9057

environment distortion have an enormous impact on the performance of SER. So, it is important to explore new strategies that can learn stable features which are invariant to nuisance factors, while maintaining discriminative information with respect to the task of emotion recognition. Recently, deep learning has been successfully applied in various fields such as speech processing (Lee *et al.*, 2009; Dahl *et al.*, 2012; Hinton *et al.*, 2012) and image understanding (Sun *et al.*, 2013; Chan *et al.*, 2014). For example, Lee *et al.* (2009) first applied convolutional deep belief networks (DBNs) to audio data and found that these learned feature representations have good performance for various audio classification tasks. Chan *et al.* (2014) proposed a simple network called PCANet and found that it is on par with the state-of-the-art features for many tasks such as texture classification and object recognition. Deep learning can also be called 'representation learning' or 'unsupervised feature learning', and it addresses mainly the problems of 'what makes better representations' and 'how to learn them'. In this paper, we want to use unsupervised feature learning technology to automatically learn emotion-related features from raw speech data. We hope that these algorithms can bring some new angles to understand human emotions. Specifically, this paper has the following contributions:

1. We apply several unsupervised feature learning methods, including the sparse auto-encoder, sparse restricted Boltzmann machines (SRBMs), and K-means clustering, to discover emotion-related features for SER with unlabeled data.

2. We present a detailed analysis of model selection with discussion on the changes of the content window size and the number of hidden layer nodes.

2 Related work

A traditional SER system consists of (1) a front-end processing unit that extracts appropriate features from the available speech data and (2) a classifier that decides the underlying emotion of speech utterance.

Various types of classifiers have been applied, such as the Gaussian mixture model (GMM) (Thapliyal and Amoli, 2012), the support vector machine (SVM) (Mao *et al.*, 2010), the hidden Markov model (HMM) (Gao *et al.*, 2012), artificial neural networks (ANNs) (Koolagudi *et al.*, 2012), k-nearest neighbor (k-NN) (Feraru and Zbancioc, 2013), and other hybrid classifiers like that proposed by Li *et al.* (2013). It seems that these classifiers have their own advantages and we can choose one according to a specific scenario. The literature does not reach a consensus about which classifier is the most suitable for SER. In this study, we focus on front-end processing for feature extraction and choose a linear SVM to make the classification.

Speech features can be grouped into four categories: (1) acoustic features, (2) linguistic features, (3) context information, and (4) hybrid features which combine acoustic features with other information sources. The most frequently used are acoustic features, which can be further classified into four types: continuous features, qualitative features, spectral features, and Teager energy operator (TEO) based features. The other features like linguistic features and context knowledge are generally combined with acoustic features to boost performance. For example, Wu *et al.* (2011) combined modulation spectral features (MSFs) with prosodic features and showed a substantial improvement. Sun and Moore (2011) reported the performance of glottal waveform parameters and TEO in classification of binary classes of four emotion dimensions, and found that TEO and the glottal parameters are on par with prosodic, spectral, and other voicing related features. Mencattini *et al.* (2014) used speech amplitude modulation with other standard features like pitch contour for SER. Koolagudi *et al.* (2012) examined pitch-related features in the Berlin emotion speech corpus. Mao *et al.* (2013) proposed a fusion algorithm which combines functional paralanguage features with accompanying paralanguage features, and evaluated its usefulness for SER. Wu and Liang (2011) found that combining acoustic-prosodic information with semantic labels can achieve higher performance than considering acoustic-prosodic information or semantic labels alone. Hybrid features combining acoustic features with linguistic and context features were proposed by Ramakrishnan and El Emary (2013). Although many speech emotion features have been explored for SER, researchers have not identified the best speech features for this task, and it is unclear whether these hand-designed features can sufficiently and efficiently characterize the emotional content of speech, due to the tight

coupling of speech emotion and other factors such as speaker and other environment distortions. The research may be more likely to combine more types of features. Moreover, the automatic extraction of emotional speech features is challenging. Most of these features are typically extracted manually or directly from transcripts.

Recently, unsupervised feature learning has been successfully applied in emotion recognition (Schmidt and Kim, 2011; Stuhlsatz *et al.*, 2011; Kim *et al.*, 2013; Le and Provost, 2013). Schmidt and Kim (2011) used DBNs to learn high-level features directly from magnitude spectra for music emotion recognition and found a subtle improvement compared with sophisticated hand-crafted features. Stuhlsatz *et al.* (2011) proposed a generalized discriminant analysis (GerDA) based on deep neural networks (DNNs) to learn compact discriminative features and achieved better performance in both unweighted and weighted recall on multiple emotion corpora. Kim *et al.* (2013) used DBNs to capture non-linear feature interactions in multimodal data and showed improvement over baseline without deep learning methods. Le and Provost (2013) proposed and evaluated a hybrid classifier based on a DBN-HMM and achieved results on a spontaneous emotion corpus.

As far as we know, the above deep learning methods use hand-crafted features as their input, except for that proposed by Schmidt and Kim (2011). We use the magnitude spectra of speech emotional data as the input to our model. However, Schmidt and Kim (2011) did not give more details about the effect these hyper-parameters may have on SER by using unsupervised feature learning algorithms. Thus, the current work may provide some insights to better understand emotion and improve recognition performance.

3 Unsupervised feature learning for speech emotion recognition

3.1 System architecture

Different from traditional feature extraction methods, we use unlabeled data to learn emotion-related feature extractors. We then train a linear SVM using training data whose features are extracted by the learned feature extractors. The system pipeline is shown in Fig. 1. The basic routine is as follows: after preprocessing using principal component analysis (PCA) and whitening, we obtain many patches from the unlabeled training data. Then we use these patches to train three unsupervised feature learning models, K-means, the sparse auto-encoder, and SRBMs. Afterwards, we use the trained filters or feature map functions to extract emotional features for audio samples. Finally, we calculate simple summary statistics (here, we use the sum) for the features learned by these three algorithms and transport them to SVM for classification.

3.2 Unsupervised feature learning algorithm

An unsupervised algorithm can be viewed as a map function $g\colon \boldsymbol{X} \to \boldsymbol{Y}$, which means that it takes the vector $\boldsymbol{X} \in \mathbb{R}^{D_x}$ as the input and outputs a new feature vector $\boldsymbol{Y} \in \mathbb{R}^{D_y}$. In the following, we use the above mentioned three unsupervised learning methods:

1. K-means clustering: K-means is one of the simplest unsupervised learning algorithms that solve the clustering problem. The procedure follows a simple way to classify a given data set through a certain number of clusters fixed a priori. The main idea is to define k centroids, one for each cluster. Then take each point belonging to a given data set and associate it to the nearest centroid. Afterwards, re-calculate the new k centroids and repeat the above steps until the centroids no longer move. The objective function of K-means is as follows:

$$J = \sum_{i=1}^{m} ||\boldsymbol{x}^{(i)} - \boldsymbol{u}_{c^{(i)}}||_2^2, \tag{1}$$

where $c^{(i)} = 1, 2, \ldots, k$. $\boldsymbol{u}_{c^{(i)}}$ is the mean of the elements of cluster $c^{(i)}$; in other words, $\boldsymbol{u}_{c^{(i)}}$ is the position with respect to centroids $c^{(i)}$. Generally speaking, when centroids are converged, J will achieve its minimal value. Once the centroids have been learned, we consider a non-linear feature mapping approach as adopted in Coates *et al.* (2011), which attempts to obtain a 'softer' representation for the input:

$$g_k(x) = \max\{0, m(\boldsymbol{d}) - d_k\}, \tag{2}$$

where $d_k = ||\boldsymbol{x} - \boldsymbol{u}_{c^{(k)}}||_2$, and $m(\boldsymbol{d})$ is the mean of the elements of $\boldsymbol{d}$. This feature mapping function outputs 0 when the distance to the centroid is above average for any feature g_k. Actually, this means that

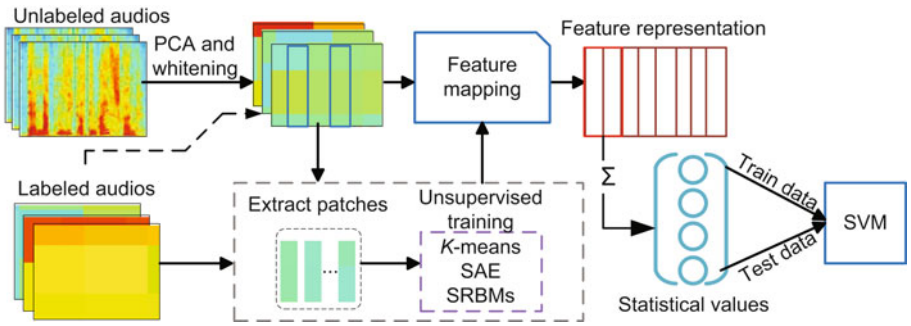


Fig. 1 System pipeline (SAE: sparse auto-encoder; SRBMs: sparse restricted Boltzmann machines)

roughly half of the features will be set to 0. So, we can see that this feature mapping approach still keeps sparsity to some extent.

2. Sparse auto-encoder: A sparse auto-encoder is a model based on an auto-encoder with an additional penalty term that encourages the hidden units to maintain a low average activation. The algorithm outputs weights $\boldsymbol{W} \in \mathbb{R}^{D_x \times D_y}$ and biases $\boldsymbol{b} \in \mathbb{R}^{D_y}$ when the following objective function is minimized:

$$J = ||f(\boldsymbol{W}^{\mathrm{T}} f(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b})+\boldsymbol{b}^{\mathrm{T}})-\boldsymbol{x}||^2+\lambda\sum_{j=1}^{K}\mathrm{KL}(\rho||\hat{\rho}_j), \tag{3}$$

$$\mathrm{KL}(\rho||\hat{\rho}_j) = \rho\log_2\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log_2\frac{1-\rho}{1-\hat{\rho}_j}, \tag{4}$$

$$\hat{\rho}_j = \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{W}_j\boldsymbol{x}_i + b_j), \tag{5}$$

where $f(t) = \mathrm{sigmoid}(t) = (1+\mathrm{e}^{-t})^{-1}$, $\mathrm{KL}(\rho||\hat{\rho}_j)$ means the Kullback-Leibler divergence between ρ and $\hat{\rho}_j$, and ρ_j, $\boldsymbol{W}_j$, and b_j are the average activation, weight, and bias of hidden node j, respectively. The parameter λ controls the relative importance of the two terms. The feature map function is defined as $g(x) = f(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b})$. The sparse auto-encoder can learn compact features when the number of hidden nodes is less than the dimensionality of the input. The hidden layer nodes above can also be used as input to the next layer in order to build a DNN.

3. Sparse restricted Boltzmann machine (SRBM): The RBMs are probabilistic graphical models that can be interpreted as stochastic neural networks. SRBM is based on RBMs with the same type of sparsity penalty as the auto-encoders. The weights $\boldsymbol{W}$ and biases $\boldsymbol{b}$ of learned SRBMs can be used for the feature mapping function, which is similarly defined as a sparse auto-encoder. In SRBM, we treat the input $\boldsymbol{x}$ as symbol $\boldsymbol{v}$, represented as visual layer nodes, and the hidden layer nodes are represented by symbol $\boldsymbol{h}$. The error function of SRBM is shown as follows:

$$\begin{aligned} J = &-\sum_{i=1}^{N}\log_2\sum_{h}P(\boldsymbol{v}^{(i)},\boldsymbol{h}^{(i)}) \\ &+\lambda\sum_{j=1}^{K}\left|\rho-\frac{1}{N}\sum_{i=1}^{N}E(h_j^{(i)}|\boldsymbol{v}^{(i)})\right|^2, \end{aligned} \tag{6}$$

where $E(\cdot)$ is the conditional expectation given the data, λ is a regularization constant, and ρ is a constant that controls the sparseness of the hidden units h_j. SRBM can be trained using contrastive divergence, which is primarily different from the training method of a sparse auto-encoder.

3.3 Feature extraction

From the above, we now obtain three feature map functions which can be used for transforming an input $\boldsymbol{X} \in \mathbb{R}^{D_x}$ to a new representation $\boldsymbol{Y} \in \mathbb{R}^{D_y}$. We apply these functions g to the input spectrogram convolutionally as the convolution operator has been shown successfully to improve performance in computer vision (Ranzato *et al.*, 2007; Chan *et al.*, 2014; Razavian *et al.*, 2014) and audio processing (Lee *et al.*, 2009; Abdel-Hamid *et al.*, 2012). Specifically, given an audio spectrogram $\boldsymbol{X}_{s_i} \in \mathbb{R}^{l\times s_i}$ $(i = 1, 2, \cdots, N)$, where N is the total number of audio samples, the indicator s_i means that the time-varying audios have different frame sizes. Note that we use the patches $\boldsymbol{X}_p \in \mathbb{R}^{l\times s_p}$ (namely, $\boldsymbol{X} \in \mathbb{R}^{D_x}$) to train the above mentioned three unsupervised feature learning algorithms rather than using $\boldsymbol{X}_{s_i}$. The symbol s_p is patch size, which we can also call the 'content window size'. In the experiments, we will discuss the optimal values of the content window size.

Throughout this paper we use the terms 'patch size' and 'content window size' interchangeably.

Given any $l \times s_p$ spectrogram patch, we can compute a new representation $\boldsymbol{y} \in \mathbb{R}^{D_y}$ for that patch through function g. Then we can compute a representation of the entire spectrogram i by applying the function g to many sub-patches. Specifically, a spectrogram $\boldsymbol{X}_{s_i}$ $(l \times s_i)$ can be split by $s_i - s_p + 1$ sub-patches when the step size is 1. Then we compute the representation $\boldsymbol{y}$ for each sub-patch of the spectrogram. To keep the same size and reduce the dimension of representations of different audio inputs, we adopt a strategy which can be understood as a simple way of pooling. Namely, we split the new representation of one spectrogram into four equal-sized blocks and compute the sum of each block. This ultimately yields a total of $4D_y$ features that can be used for classification.

4 Experiments and analysis

4.1 Datasets and experimental setup

We analyze the performance of the proposed unsupervised feature learning algorithms on three public available databases: the Surrey Audio-Visual Expressed Emotion (SAVEE) database (Haq and Jackson, 2009), the Berlin Emotional Speech Database (Emo-DB) (Burkhardt *et al.*, 2005), and the eNTERFACE'05 emotion database (Martin *et al.*, 2006). The SAVEE database consists of recordings from 4 male actors in 7 different emotions, in total 480 British English utterances with 6 basic emotions (anger, disgust, fear, happiness, sadness, and surprise) and neutral. Emo-DB consists of 535 utterances recorded in German with 10 actors. The corpus also covers 7 emotional states (anger, disgust, boredom, joy, fear, sadness, and neutral). The last eNTERFACE'05 emotion corpus contains 42 subjects with 1293 utterances recorded in English. It consists of induced anger, disgust, fear, joy, sadness, and surprise speaker emotions. Each subject was instructed to listen to 6 successive short stories, each of them eliciting a particular emotion.

First, we evaluate the effects of the two important parameters, i.e., the number of hidden nodes and content window size, using cross-validation on the eNTERFACE database. Then, we report the results achieved on the other two emotional databases using the parameter setting that our analysis suggests is best overall. At the same time, we stack a second layer using the first hidden layer nodes as input to the sparse auto-encoder and SRBMs. To clarify, we use SAE L.1 and SAE L.2 to represent that the network has one hidden layer and two hidden layers for the sparse auto-encoder, respectively. SRBMs L.1 and SRBMs L.2 apparently have a similar meaning for SRBMs.

We set a fixed size of the patch and choose a specific number of hidden nodes, and then we determine the final number of hidden nodes based on the accuracy. Afterwards, we can evaluate the patch size with the former defined number of hidden nodes. Specifically, we choose 30 subjects of the eNTERFACE database as the unlabeled data to train the three unsupervised feature learning algorithms. Then we extract features of the remaining 14 subjects using the former learned feature extractors and apply them to a linear SVM classifier using five-fold cross-validation. Similarly, we choose 188 unlabeled, 347 labeled speech data for Emo-DB and 240 unlabeled, 240 labeled speech data for the SAVEE corpus. Note that the subjects of unlabeled and labeled speech data of three databases are different from each other.

The basic pre-processing procedure is as follows. We first convert the time-domain signals into spectrograms. The spectrogram has a 20 ms window size with a 10 ms overlap. The spectrogram can be further processed using PCA and whitening (with 80 components) to reduce its dimensionality. We randomly pick up 10 000 patches of unlabeled data of all three databases, which are used for unsupervised feature learning. In our experiments, we use the same setting for all learning algorithms.

4.2 Visualization

To show what the three unsupervised feature learning algorithms have really learned from the speech data, we first adopt the strategy used by Lee *et al.* (2009) to visualize some randomly selected weights (or centroids) learned by the sparse auto-encoder, SRBMs, and K-means models. Fig. 2 shows that the weights (or centroids) are dependent upon energy in specific frequency bands. So, we can infer that the weights (or centroids) transform the spectrogram into a common feature space, where it can be more powerful to represent emotional speech signals.

In the field of computer vision, the auto-encoders and RBMs yield localized filters that resemble Gabor filters. The result of the visualization of K-means looks more orderly than those of the other two frequently used deep learning methods.

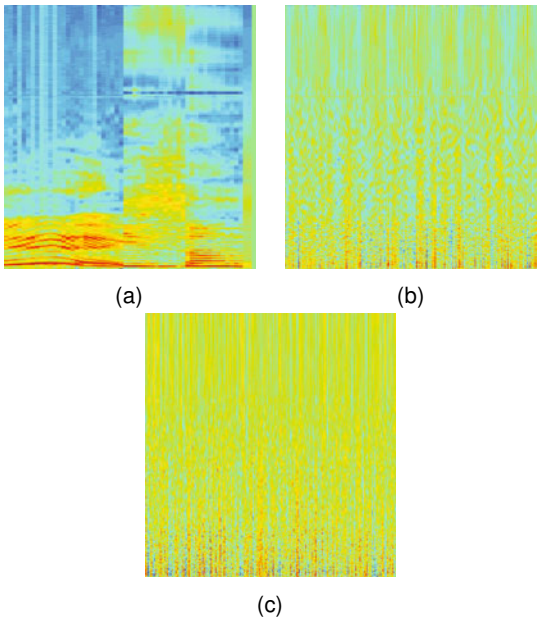


Fig. 2 Randomly selected bases (or centroids) trained on the eNTERFACE database using different learning algorithms: (a) K-means; (b) SAE; (c) SRBM

Secondly, we investigate reconstructing the original spectrogram back from the transformed features to further show the properties of our feature learning method. For the sparse auto-encoder and SRBMs, we just take the reconstruction values after they have been learned, but it is hard to obtain the input reconstruction for K-means since we have adopted a 'soft' way to extract features. So, we use the learned features of the specific input instead for K-means.

In Fig. 3, the heatmaps for the reconstructed spectra of the sparse auto-encoder and SRBMs may exhibit a property of sparsity. The reconstruction also reveals that the learned weights (Fig. 2) are dependent upon energy in specific frequency bands to some extent, which is also shown by DBNs in Schmidt and Kim (2011). For the K-means learned features, we also observe highly sparse features which must be due to the 'soft' strategy that we have used.

4.3 Number of hidden nodes

To evaluate the effect of the number of hidden nodes, we first choose a patch size (or the content window size) used in Mohamed *et al.* (2012). Then we consider feature representations with 50, 100, 200, 400, 600, and 800 hidden nodes or centroids. Fig. 4 shows the effect of increasing the number of hidden nodes: all algorithms generally achieve high performance with more hidden nodes, as expected. From Fig. 4, when the number of hidden nodes is set to 800, the performance drops a little. The reason may be that 600 hidden nodes may already have excellent representational power. The performances of the sparse auto-encoder and SRBMs are somewhat close to each other. Unlike Coates *et al.* (2011), we do not find that K-means can achieve higher performance than the sparse auto-encoder or SRBMs. We use 600 hidden nodes to evaluate the effect of the content window size.

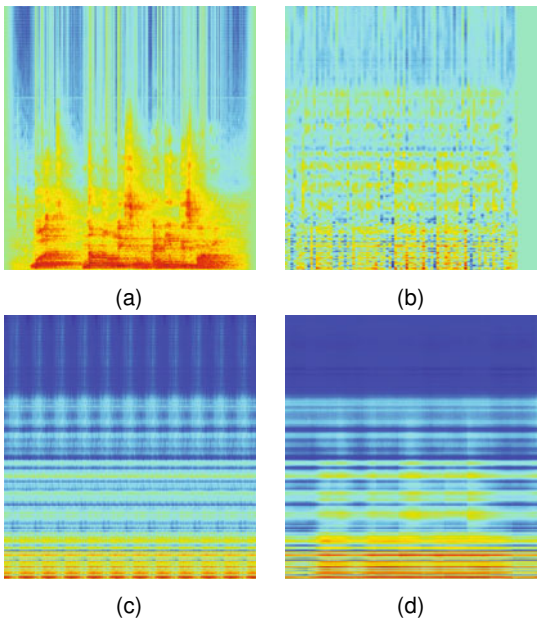


Fig. 3 Spectra reconstruction and learned features: (a) log view of magnitude of the common spectra input; (b) log view of features learned by K-means; (c) log view of SAE magnitude spectra reconstruction; (d) log view of SRBM magnitude spectra reconstruction

4.4 Effect of the content window size

We consider three different content window sizes, 7, 17, and 27. Fig. 5 clearly shows the effect of the patch size: all algorithms achieve higher performance using the larger patch. The reason may be that the emotion spans a long content on the speech. Our results happen to coincide with those in Le and Provost (2013), which reached a similar conclusion.

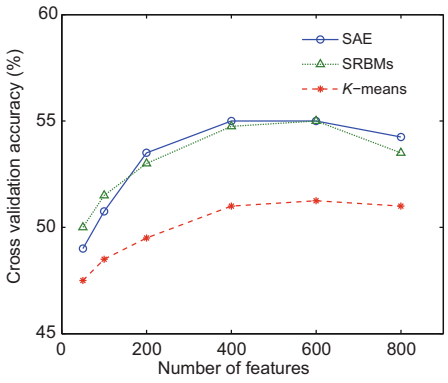


Fig. 4 Effect of the number of hidden nodes

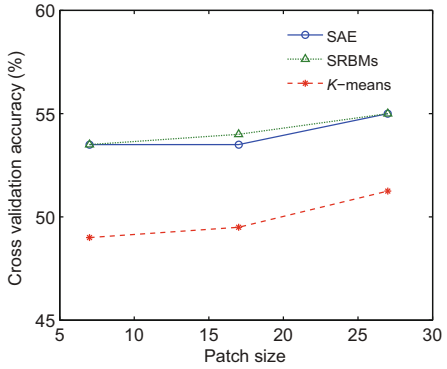


Fig. 5 Effect of the content window size

Generally speaking, a larger content window size allows us to recognize more complex features that cover a large region of the speech signals. However, a larger patch size may lead to a need for more data and to learn more features. This may be challenging in limited circumstances. In the following, we use 600 hidden nodes and patch size 27 to evaluate the performances on the other two speech emotional databases.

4.5 Performance evaluation

Using the above determined parameters (600 hidden nodes and a content window size of 27), we evaluate the performances on the other two emotional databases, Emo-DB and SAVEE. As mentioned above, we stack a second layer using the first hidden layer nodes as input to the sparse autoencoder and SRBMs to construct a second layer. The number of the second hidden layer nodes is also set to 600. We also use the spectrogram as the raw features without learning to show that these learning algorithms indeed work. We use simple summary statistics for each frequency bin over the spectrogram. Just like the way of 'pooling' described in Section 3.3, we split the specific spectrogram into four equal-sized blocks and then compute the sum of each block. Finally, we combine the four blocks as features to train the classifier.

The final classification results with these settings are reported in Table 1. The results of all experiments are obtained by five-fold cross-validation. As shown in Table 1, the learned features (regardless of the number of hidden layers or the methods used) apparently outperform raw spectrogram features for Emo-DB and eNTERFACE. However, the accuracy of classification of the raw features is on a par with the learned features for SAVEE. The reason may be that the circumstances of Emo-DB or eNTERFACE are more challenging than the SAVEE corpus. For example, the number of subjects of Emo-DB or eNTERFACE is larger than that of the SAVEE corpus. In fact, the number of subjects of SAVEE for doing classification is just 2. On the other hand, the three unsupervised feature learning algorithms may produce features which are robust to speaker variation or other factors for Emo-DB and eNTERFACE. In addition, Table 1 shows that the two-layer network cannot explicitly improve performance compared to a single layer since a single-layer network may already have a good representational power under our experimental conditions.

Table 1 Final accuracy on three public databases

Method	Accuracy (%)		
	SAVEE	Emo-DB	eNTERFACE
RAW	86.67	22.48	22.25
K-means	85.83	71.49	51.25
SAE L.1	87.50	65.12	55.00
SAE L.2	86.66	67.43	55.00
SRBMs L.1	86.66	71.45	55.50
SRBMs L.2	85.42	71.16	56.00

5 Conclusions

In this paper, we propose three unsupervised learning algorithms to learn emotion-related features for SER. We evaluate the effects of the two important parameters, number of hidden nodes and content window size, on the eNTERFACE database with these algorithms, and show that a larger content window and more hidden nodes can contribute to better performance. Our work may provide insights to help understand emotion and improve recognition

technologies.

In the future, we will consider using labeled speech data to fine-tune the parameters of the network to further improve the performance. Also, it is necessary to search or assemble large emotional data to prepare enough data for a deep network. In another direction, we will seek salient features which can be robust to environmental distortion or speaker variation by adding some penalty terms, since a deep learning method may be sensitive to small perturbations in the input features.

References

Abdel-Hamid, O., Mohamed, A.R., Jiang, H., *et al.*, 2012. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.4277-4280. [doi:10.1109/ICASSP.2012.6288864]

Burkhardt, F., Paeschke, A., Rolfes, M., *et al.*, 2005. A database of German emotional speech. Interspeech, p.1517-1520.

Chan, T.H., Jia, K., Gao, S., *et al.*, 2014. PCANet: a simple deep learning baseline for image classification? arXiv preprint, arXiv:1404.3606.

Coates, A., Ng, A.Y., Lee, H., 2011. An analysis of single-layer networks in unsupervised feature learning. Int. Conf. on Artificial Intelligence and Statistics, p.215-223.

Dahl, G.E., Yu, D., Deng, L., *et al.*, 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio Speech Lang. Process.*, **20**(1):30-42. [doi:10.1109/TASL.2011.2134090]

El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: features, classification schemes, and databases. *Patt. Recogn.*, **44**(3):572-587. [doi:10.1016/j.patcog.2010.09.020]

Feraru, M., Zbancioc, M., 2013. Speech emotion recognition for SROL database using weighted KNN algorithm. Int. Conf. on Electronics, Computers and Artificial Intelligence, p.1-4. [doi:10.1109/ECAI.2013.6636198]

Gao, H., Chen, S.G., An, P., *et al.*, 2012. Emotion recognition of Mandarin speech for different speech corpora based on nonlinear features. IEEE 11th Int. Conf. on Signal Processing, p.567-570. [doi:10.1109/ICoSP.2012.6491552]

Gunes, H., Schuller, B., 2013. Categorical and dimensional affect analysis in continuous input: current trends and future directions. *Image Vis. Comput.*, **31**(2):120-136. [doi:10.1016/j.imavis.2012.06.016]

Haq, S., Jackson, P.J., 2009. Speaker-dependent audio-visual emotion recognition. Auditory-Visual Speech Processing, p.53-58.

Hinton, G., Deng, L., Yu, D., *et al.*, 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.*, **29**(6):82-97. [doi:10.1109/MSP.2012.2205597]

Kim, Y., Lee, H., Provost, E.M., 2013. Deep learning for robust feature generation in audiovisual emotion recognition. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.3687-3691. [doi:10.1109/ICASSP.2013.6638346]

Koolagudi, S.G., Devliyal, S., Barthwal, A., *et al.*, 2012. Emotion recognition from semi natural speech using artificial neural networks and excitation source features. *In*: Contemporary Computing. Springer Berlin Heidelberg, p.273-282.

Le, D., Provost, E.M., 2013. Emotion recognition from spontaneous speech using hidden Markov models with deep belief networks. IEEE Workshop on Automatic Speech Recognition and Understanding, p.216-221. [doi:10.1109/ASRU.2013.6707732]

Lee, H., Pham, P., Largman, Y., *et al.*, 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. Advances in Neural Information Processing Systems, p.1096-1104.

Li, L., Zhao, Y., Jiang, D., *et al.*, 2013. Hybrid deep neural network–hidden Markov model (DNN-HMM) based speech emotion recognition. Humaine Association Conf. on Affective Computing and Intelligent Interaction, p.312-317. [doi:10.1109/ACII.2013.58]

Mao, Q., Wang, X., Zhan, Y., 2010. Speech emotion recognition method based on improved decision tree and layered feature selection. *Int. J. Human. Robot.*, **7**(2):245-261. [doi:10.1142/S0219843610002088]

Mao, Q.R., Zhao, X.L., Huang, Z.W., *et al.*, 2013. Speaker-independent speech emotion recognition by fusion of functional and accompanying paralanguage features. *J. Zhejiang Univ.-Sci. C (Comput. & Electron.)*, **14**(7):573-582. [doi:10.1631/jzus.CIDE1310]

Martin, O., Kotsia, I., Macq, B., *et al.*, 2006. The eNTERFACE'05 audio-visual emotion database. Proc. Int. Conf. on Data Engineering Workshops, p.8. [doi:10.1109/ICDEW.2006.145]

Mencattini, A., Martinelli, E., Costantini, G., *et al.*, 2014. Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. *Knowl.-Based Syst.*, **63**:68-81. [doi:10.1016/j.knosys.2014.03.019]

Mohamed, A.R., Dahl, G.E., Hinton, G., 2012. Acoustic modeling using deep belief networks. *IEEE Trans. Audio Speech Lang. Process.*, **20**(1):14-22. [doi:10.1109/TASL.2011.2109382]

Nicolaou, M.A., Gunes, H., Pantic, M., 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affect. Comput.*, **2**(2):92-105. [doi:10.1109/T-AFFC.2011.9]

Pantic, M., Nijholt, A., Pentland, A., *et al.*, 2008. Human-centred intelligent human? Computer interaction (HCI2): how far are we from attaining it? *Int. J. Auton. Adapt. Commun. Syst.*, **1**(2):168-187. [doi:10.1504/IJAACS.2008.019799]

Ramakrishnan, S., El Emary, I.M., 2013. Speech emotion recognition approaches in human computer interaction. *Telecommun. Syst.*, **52**(3):1467-1478. [doi:10.1007/s11235-011-9624-z]

Ranzato, M., Huang, F.J., Boureau, Y.L., *et al.*, 2007. Unsupervised learning of invariant feature hierarchies with applications to object recognition. IEEE Conf. on Computer Vision and Pattern Recognition, p.1-8. [doi:10.1109/CVPR.2007.383157]

Razavian, A.S., Azizpour, H., Sullivan, J., *et al.*, 2014. CNN features off-the-shelf: an astounding baseline for recognition. arXiv preprint, arXiv:1403.6382.

Schmidt, E.M., Kim, Y.E., 2011. Learning emotion-based acoustic features with deep belief networks. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, p.65-68. [doi:10.1109/ASPAA.2011.6082328]

Stuhlsatz, A., Meyer, C., Eyben, F., *et al.*, 2011. Deep neural networks for acoustic emotion recognition: raising the benchmarks. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.5688-5691. [doi:10.1109/ICASSP.2011.5947651]

Sun, R., Moore, E.II, 2011. Investigating glottal parameters and Teager energy operators in emotion recognition. *LNCS*, **6975**:425-434. [doi:10.1007/978-3-642-24571-8_54]

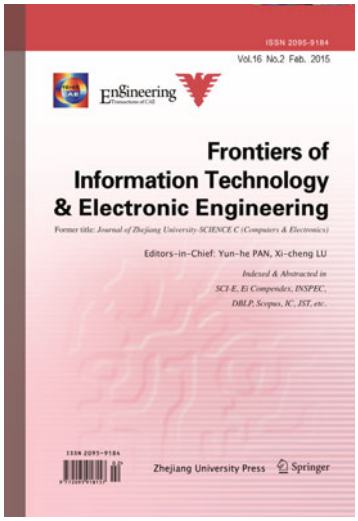
Sun, Y., Wang, X., Tang, X., 2013. Deep learning face representation from predicting 10,000 classes. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, p.1891-1898. [doi:10.1109/CVPR.2014.244]

Thapliyal, N., Amoli, G., 2012. Speech based emotion recognition with Gaussian mixture model. *Int. J. Adv. Res. Comput. Eng. Technol.*, **1**(5):65-69.

Wu, C.H., Liang, W.B., 2011. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Trans. Affect. Comput.*, **2**(1):10-21. [doi:10.1109/T-AFFC.2010.16]

Wu, S., Falk, T.H., Chan, W.Y., 2011. Automatic speech emotion recognition using modulation spectral features. *Speech Commun.*, **53**(5):768-785. [doi:10.1016/j.specom.2010.08.013]

FITEE: Call for papers



Editors-in-Chief: Yun-he Pan, Xi-cheng Lu

Frontiers of Information Technology & Electronic Engineering (ISSN 2095-9184, monthly), *FITEE* for short, is an international peer-reviewed journal launched by Chinese Academy of Engineering (CAE) and Zhejiang University, co-published by Springer & Zhejiang University Press. *FITEE* is aimed to publish the latest implementation of applications, principles, and algorithms in the broad area of Electrical and Electronic Engineering, including but not limited to Computer Engineering, Telecommunications, Control Systems, Robotics, Radio Engineering, Signal Processing, Power Engineering, Systems Engineering, Electronics, Microelectronics, etc.

FITEE is formerly known as *Journal of Zhejiang University-SCIENCE C (Computers & Electronics)* (2010–2014), which has been covered by SCI-E since 2010. Authors of manuscripts submitted or accepted come from 40+ countries and regions, including mainland China, Taiwan, Malaysia, Iran, Korea, Spain, Germany, UK, Greece, USA, Brazil, etc. There are different types of articles for your choice, including **research articles, review articles, science letters, perspective, new technical notes and methods**, etc.

Highlights (metrics & services):

- Key metrics:
  Impact factor: 0.38
  Peer review period: 1–3 months
  From submission to publication (currently): <10 months
  Frequency of publication: monthly
  Editorial board: 16 foreign members, 29 domestic members (including 16 members of CAE)
- Timely and high-quality service for authors and readers
- Rigorous editing and proof-reading
- **Article in press**: Accepted articles will be pushed online immediately after the acceptance
- Innovative techniques adopted:
  CrossMark, to track content changes
  ORCID, to connect research and researchers
- **Peer reviewer comments** (before publication) are selected by editor to be demonstrated and **open peer comments** (after publication) can be provided by readers on the article page
- **English summary** is provided for each paper to give readers a quick view and **Chinese summary** to a wider audience of Chinese readers, both freely accessible
- Abstracted/Indexed in: SCI-E, EI-Compendex, SCOPUS, INSPEC, Google Scholar, DBLP, etc.
- Full text is available from www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com

Thanks for your attention and welcome your contribution!

Online submission:
http://www.editorialmanager.com/zusc/

Manuscript guidelines:
http://www.zju.edu.cn/jzus/manuscript.php

Contact:
Editorial Office of *J. Zhejiang Univ.-SCIENCE (A/B)* & *FITEE*
38 Zheda Road, Hangzhou 310027, China
Managing Editors: Helen Zhang & Ziyang Zhai
jzus@zju.edu.cn; jzus_zzy@zju.edu.cn
+86-571-87952276/87952783