Zheng-wei Huang, Wen-tao Xue, Qi-rong Mao, 2015. Speech emotion recognition with unsupervised feature learning. *Froniters of Information Technology & Electronic Engineering*, **16**(5):358-366. [doi:10.1631/FITEE.1400323]

Speech emotion recognition with unsupervised feature learning

Key words: Speech emotion recognition, Unsupervised feature learning, Neural network, Affect computing

Corresponding author: Qi-rong Mao
E-mail: mao_qr@mail.ujs.edu.cn
ORCID: http://orcid.org/0000-0002-5021-9057

Motivation

- Much of the actual effort in deploying systems of speech emotion recognition (SER) goes into the design of an appropriate representation of speech signals.
- Current research on feature extraction:
  - Most of these features are typically extracted mannually or directly from transcripts.
  - The researchers have not identified the best speech features for SER.
  - It is unclear whether these hand-designed features can sufficiently and efficiently characterize the emotional content of speech.

Our work

– Apply several unsupervised feature learning methods, including the sparse auto-encoder (SAE), sparse restricted Boltzmann machines (SRBMs), and *K*-means clustering, to discover emotion-related features for SER with unlabeled original speech signals.

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

System pipeline

Our feature learning method can be divided into two parts (Fig.1):

1. Using three unsupervised learning algorithms (*K*-means, SAE, SRBMs) to learn feature mapping functions which can be used for extracting emotional features of speech signals.

After preprocessing using principal component analysis (PCA) and whitening, we extract many patches from the unlabeled training data.

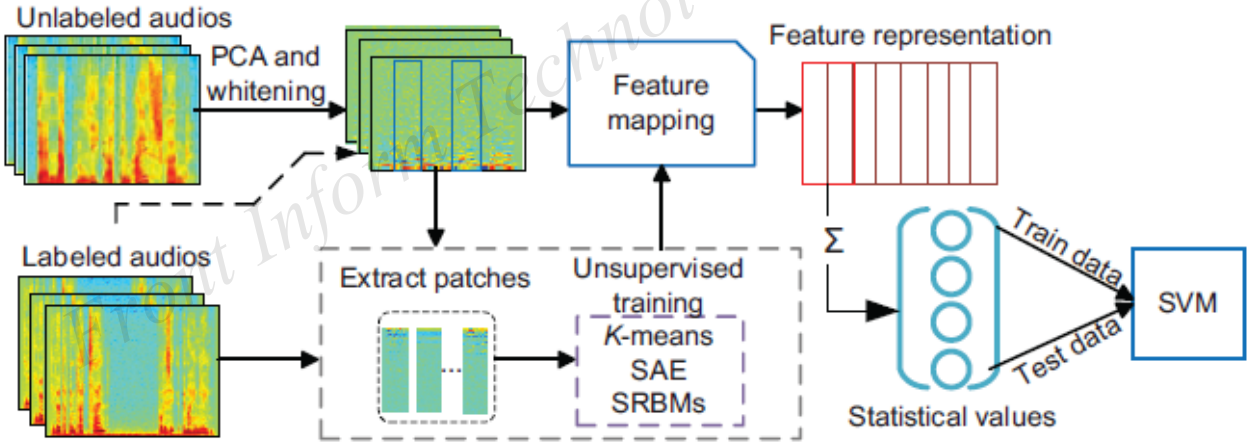
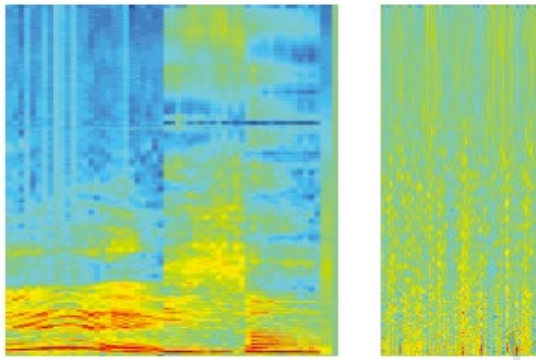


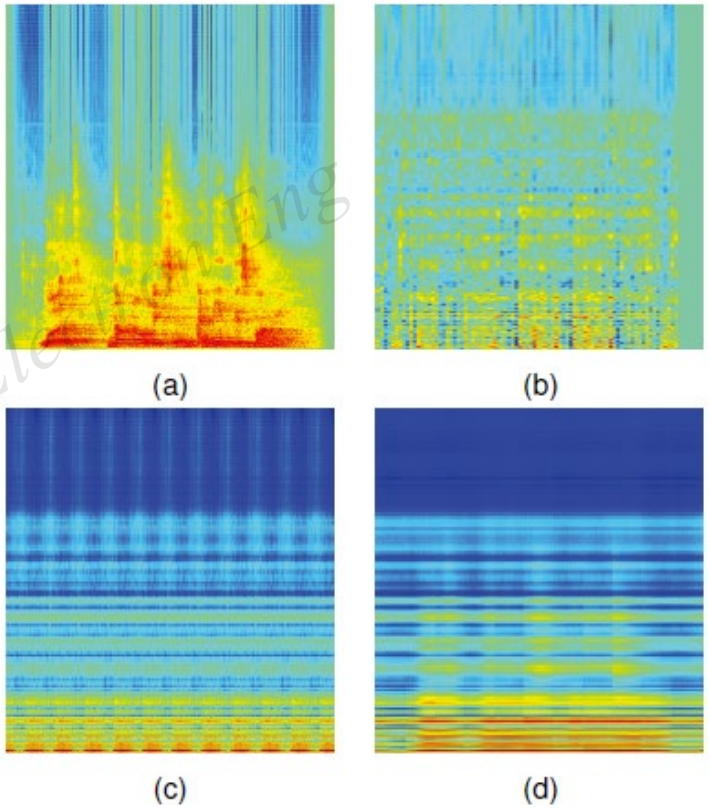
Fig. 1 System pipeline (SAE: sparse auto-encoder; SRBMs: sparse restricted Boltzmann machines)

2. Train a linear SVM and make classification.

Visualization

Randomly selected bases (or centroids) trained on the eNTERFACE'05 database using different learning algorithms: (a) *K*-means; (b) SAE; (c) SRBM





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

Hyperparameter selection

- Set a fixed size of the patch and choose a specific number of hidden nodes (consider 50, 100, 200, 400, 600, 800, Fig.4);
- Choose the patch size (consider 7, 17, 27) with the former defined number of hidden nodes (Fig.5).
- Use 600 hidden nodes and patch size 27 to evaluate the performances on the three public emotional databases.

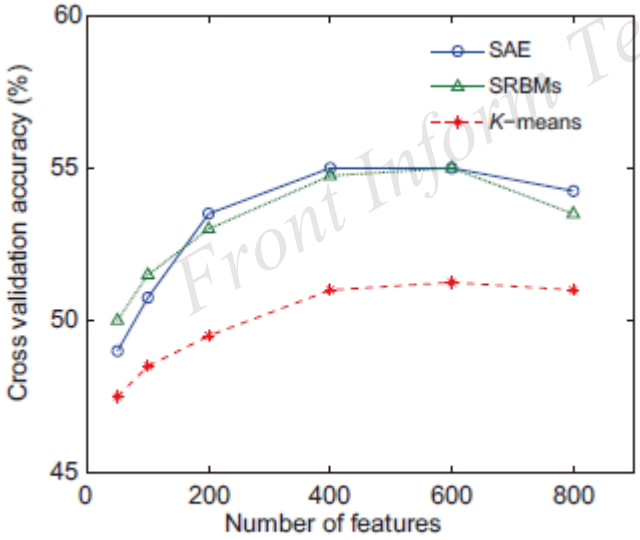


Fig. 4 Effect of the number of hidden nodes

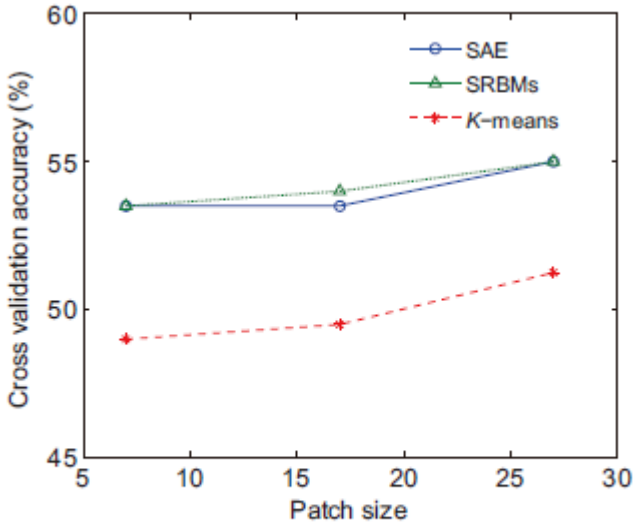


Fig. 5 Effect of the content window size

Major results

- ✓ Three emotional databases: Emo-DB, SAVEE, eNTERFACE
- ✓ RAW: the original spectrogram representations
- ✓ *K*-means: feature representations learned by *K*-means
- ✓ SAE L.1, SAE L.2: feature representations learned by single layer and two layers with sparse auto-enocder respectively
- ✓ SRBM L.1, SRBM L.2: feature representations learned by single layer and two layers with sparse restricted boltzmann machines respectively
- ✓ Results (Table 1)

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

Conclusions

- The three unsupervised learning methods can produce features which are sparse and robust to speaker variation or other distortions.
- A larger content window size and more hidden nodes can contribute to better performance.
- Compared to raw features, the learned features obviously boost the performance of SER.