Locality Preserved Joint Nonnegative Matrix Factorization for Speech Emotion Recognition

Seksan MATHULAPRANGSAN\textsuperscript{[a]}, Yuan-Shan LEE\textsuperscript{[b]}, Nonmembers, and Jia-Ching WANG\textsuperscript{[a],[c]}, Member

SUMMARY This study presents a joint dictionary learning approach for speech emotion recognition named locality preserved joint nonnegative matrix factorization (LP-JNMF). The learned representations are shared between the learned dictionaries and annotation matrix. Moreover, a locality penalty term is incorporated into the objective function. Thus, the system’s discriminability is further improved.

key words: NMF, joint dictionary learning, locality preserving, speech emotion recognition, information extraction

1. Introduction

An emotional state of a speaker’s speech is the hidden-important information, which is significantly useful for human-computer interaction, e.g., computer tutorial system, home health care systems, social networking, in-car board system, etc.\[1\], \[2\]. Therefore, many studies applied a number of approaches to recognize human speech emotions such as acoustic speech, facial expressions, and psychophysical data. Among the aforementioned strategies, speech emotion recognition has still been one of a challenge task [3].

Generally, the recognition process of speech emotions consists of feature extraction and classification. In the first step, prosody is the most common feature used in emotional speech processing. Instead of using prosodic information, spectral-based features such as Cepstral features, Mel-frequency Cepstral coefficients (MFCCs), linear predictive Cepstral coefficients (LPCCs), spectrogram, and Mel-spectrogram are the other well-known methods. The advantage that they can be used to refine energetic information of the speech spectrum make them can reveal the information of each emotional speech characteristics. Using spectral features for emotion recognition has indicated in previous studies [4]. In classification step, K-nearest neighbors (KNNs), hidden Markov models (HMMs) [5], Gaussian mixture models (GMMs), and support vector machines (SVMs) are the present favored speech emotion classifiers. To select the proper classifier to recognize speech emotion is crucial because it affects directly to the recognition performance.

Recently, supervised dictionary learning [6]–[8] played an important role as an effective approach to improve the performance of many recognition tasks. D-KSVD [8], a supervised dictionary learning, uses the combination of dictionary learning and classifier learning. Nonnegative matrix factorization (NMF) is another well-known dictionary learning method for many research fields including speech recognition. Komatsu et al. [9] developed an NMF method to detect acoustic events, which the NMF representations are used as features to train an SVM. Mesaros et al. [10] introduced a coupled NMF (C-NMF) that decomposes both spectral representation and the corresponding labels. The C-NMF then bypasses the supervised learning of the SVM.

To the best of our knowledge, there are no joint NMF approach that take advantage of using locality information of the acoustic features, especially emotional speech.

In this paper, we introduce a novel approach for speech emotion recognition named locality preserved joint NMF (LP-JNMF). Our hypotheses is that the locality information of acoustic signals can be used to promote the discriminative power of the joint-NMF dictionary. The main contributions are two-fold. Firstly, based on the proposed objective function, the discriminative power of the dictionary is increased and the linear classifier can be learned jointly. The recognition result can be obtained easily from the activation matrix of the test samples and the learned linear classifier. Secondly, the locality preserving term is incorporated into the objective function of the joint NMF. Consequently, the discriminability of the learned dictionary is further improved.

2. Related Works

2.1 Nonnegative Matrix Factorization

 Recently, NMF became a famous algorithm for speech emotion task [11], [12] since it can emphasize the difference between distinct emotions while de-emphasizes the diversity of linguistic meanings among various speakers [11]. Assume that the dictionary \( A \) is composed of \( I \) exemplars, then we have \( A = \{a_1, a_2, \ldots, a_I\} \in \mathbb{R}^{F \times I} \), where \( a_i \) is the \( i \)-th exemplar, and \( F \) is the basis dimension. Then, the acoustic sample at the \( l \)-th frame \( y_l \in \mathbb{R}^{F \times 1} \) can be represented as follows.

\[
y_l \approx Ah_l = \sum_{i=1}^{I} a_i h_{i,l} \quad s.t.A \succeq 0, h_l \geq 0,
\]

where \( h_{i,l} \) is the nonnegative weight of the \( i \)-th exemplar of
the activation vector $\mathbf{h}_t = [h_{1,t}, h_{2,t}, \ldots, h_{L,t}] \in \mathbb{R}^{L \times 1}$. Assume that each sample is modeled independently, the spectrogram of each utterance can be represented as follows.

$$
\mathbf{Y} \approx \mathbf{Y} = \mathbf{A} \mathbf{H},
$$

where $\mathbf{Y} \in \mathbb{R}^{F \times M}$ is the spectrogram, $M$ is the number of frames of the represented utterance, and $\mathbf{H} \in \mathbb{R}^{L \times M}$ is the corresponding activation matrix for which the column vector is an activation vector $\mathbf{h}_t$. The objective of NMF is to minimize the distance between the data $\mathbf{Y}$ and the approximation $\hat{\mathbf{Y}}$, i.e., $\mathcal{D}(\mathbf{Y}||\hat{\mathbf{Y}})$. $\mathbf{A}$ and $\mathbf{H}$ can be optimized by multiplicative updating rules based on gradient descent. Finally, the learned activation matrix $\mathbf{H}$ is used to train SVM [9].

### 2.2 Joint NMF (J-NMF)

Joint dictionary approach is another remarkable algorithm for speech and speech emotion problems. Wu et al. [13] introduced joint NMF (J-NMF) for voice conversion. This method used two types of features, multiple-frame low-resolution exemplars and single-frame high-resolution exemplars, to estimate the activation matrix simultaneously. Fu et al. [14] introduced joint dictionary learning-based NMF (JD-NMF) to improve patients’ speech intelligibility after oral surgery. The JD-NMF was designed to learn the source and the target dictionaries simultaneously from the common activation matrix.

The J-NMF objective function [10] can be presented as,

$$
\langle \mathbf{A}^{(1)}, \mathbf{A}^{(2)} \rangle = \arg \min_{\mathbf{A}^{(1)}, \mathbf{A}^{(2)} > 0} \{\eta_1 \mathcal{D}(\mathbf{Y}^{(1)}, \mathbf{A}^{(1)} \mathbf{H}) + \eta_2 \mathcal{D}(\mathbf{Y}^{(2)}, \mathbf{A}^{(2)} \mathbf{H})\},
$$

where $\mathcal{D}()$ is the Kullback-Leibler divergence, $\mathbf{Y}^{(1)} \in \mathbb{R}^{F \times M}$ is the spectrum and $\mathbf{Y}^{(2)} \in \mathbb{R}^{E \times M}$ is the class activity annotation matrix. $F$, $M$, and $E$ are the number of frequency bins, frames, and classes, respectively. $\eta_1$ and $\eta_2$ are the weights associated with divergences. This optimization problem can be solved by using generalized linear models [15]. After that, the estimated annotation matrix is obtained by $\hat{\mathbf{Y}}^{(2)} = \mathbf{A}^{(2)} \mathbf{H}_{test}$, where $\mathbf{H}_{test}$ is the minimizer of $\mathcal{D}(\mathbf{Y}^{(1)}||\hat{\mathbf{A}}^{(1)} \mathbf{H})$.

### 3. Locality Preserved NMF

#### 3.1 System Overview

Figure 1 presents the system overview of the proposed LP-JNMF, which is divided into two parts: the training step and the testing step. In the training step, speech signal inputs are extracted to be Mel-spectrograms $\mathbf{X}$, which are used to calculate their locality preserving projection $\mathbf{L}$. Then, the class annotations $\mathbf{V}$ are achieved. After that, all of them are fed into the J-NMF to decomposes a dictionary and its corresponding activation matrix $\mathbf{H}_{train}$. The dictionary composed of three parts: the feature-learned bases $\mathbf{A}$, the classifier $\mathbf{W}$, and the locality-learned bases $\mathbf{P}$, which the first two parts are shared in the testing step.

In the testing step, the test signals are extracted to be Mel-spectrograms as the testing features $\mathbf{Y}$ and fed into J-NMF along with the feature-learned bases from the training step to compute the testing activation matrix $\mathbf{H}_{test}$. Finally, the recognition result $\hat{\mathbf{V}}$ can be easily obtained by multiplying the classifier and the testing activation matrix.

#### 3.2 Locality Preserving Projection

The local structure of a distributed data point can be modeled by using a nearest neighbor graph. This geometric structure can be considered as locality information of the data, which can be used to promote the discriminability of a learned dictionary. Locality preserving projections (LPP) [16] is a method using geometric structure. LPP optimizes the neighborhood structure preserving to generate linear projective maps. The LPP objective function can be exhibitd as,

$$
\min \sum_{i} (x_i - y_i) L_{ij},
$$

where $x_i$ is the low-dimensional representation of the input signal $y_i$. The similarity matrix $\mathbf{L}$ is defined as,

$$
L_{ij} = \begin{cases} 
\exp \left( \frac{-\|y_i - y_j\|^2}{\rho} \right) &; e(y_i, y_j) = 1 \\
0 &; e(y_i, y_j) = 0
\end{cases},
$$

where $e(y_i, y_j) = 1$ for $k$-nearest $y_j$ according to $y_i$. The di-vider $\rho$ is a tunable parameter.
3.3 The Proposed LP-JNMF

Let $X \in \mathbb{R}^{F \times M}$ be the training Mel-spectrogram features, $V \in \mathbb{R}^{E \times M}$ be the class annotations, and $L \in \mathbb{R}^{F \times M}$ be the locality preserving term. The objective function of the proposed LP-JNMF can be calculated as follows.

$$\langle A, W, P \rangle = \arg \min _{A, W, P} \{ d(X, AH) + \lambda_1 d(V, WH)$$

$$+ \lambda_2 d(L, PH) \} \ s.t. \forall [a_i, w_j, p_i, h_i \geq 0],$$

(6)

where $\lambda_1$ and $\lambda_2$ are the parameters used to balance the influence among reconstruction error, annotation error, and locality preserving error. The locality-learned bases $P$ map the activation matrix $H$ into locality preserving code space. $W$ is the linear classifier. The target matrix $L$ is the locality preserving coefficient of $X$, which is calculated as,

$$L_{ij} = \begin{cases} 
\exp \left(-\frac{|x_i - x_j|^2}{\rho}\right) & ; e(x_i, x_j) = 1 \\
-(1 - \exp \left(-\frac{|x_i - x_j|^2}{\rho}\right)) & ; e(x_i, x_j) = 0 
\end{cases},$$

(7)

where $e(x_i, x_j) = 1$ for $k$-nearest $x_j$, which belongs to the same class as $x_i$, and $e(x_i, x_j) = 0$ for $k$-nearest $x_j$, which belongs to a different class of $x_i$. Practically, J-NMF can obtain the feature-learned bases $A$, the classifier $W$, and the locality-learned bases $P$ simultaneously. So that, (6) can be rewritten as,

$$\langle A, W, P \rangle = \arg \min _{A, W, P} \{ d \left( \frac{X}{A}^T, \frac{X}{A}^L \right)$$

$$- \frac{A}{W} ) H \} \ s.t. \forall [a_i, w_j, p_i, h_i \geq 0].$$

In the testing step, the corresponding activation code of a sample $y$ can be obtained as,

$$h_{test} = \arg \min _{h} d(y - Ah_{test}) \ s.t. \forall (h_{test}) \geq 0.$$ 

(9)

Finally, the recognition result is calculated as, $\hat{v} = Wh_{test}$.

4. Experiments

The experiments were conducted on the extended version of MHMC speech emotion database [17], which contains four emotional classes, i.e., angry, sad, happy, and neutral, each class has 600 samples. We compared the proposed LP-JNMF with five baseline methods including D-KSVD [8], LC-KSVD [18], LP-KSVD [19], J-NMF [10], and Gr-NMF [20]. The J-NMF and Gr-NMF are based on Kullback Leibler divergence. Moreover, there is no mixtures of local dictionaries in J-NMF and Gr-NMF, therefore, this issue is not be considered.

4.1 Effect of Different Sizes of Bases

Since the proposed LP-JNMF is a NMF-based method, which the number of bases can affect the recognition performance. In this experiment, the proper bases size of the proposed method was sought by performing 5-fold cross validation on different bases sizes, which are 10, 20, 30, 40, and 50. The iteration was fixed to 200.

The results in Table 1 indicate that the proposed LP-JNMF performs best when the number of bases is 50. Therefore, the size of bases was set to 50 in the rest experiments. Note that we also did the same way on the NMF baselines to seek for their fitting bases size used in the rest experiments.

4.2 Effect of Different Types of Features

As we mention in Sect. 1, generally the most common used feature in this task is prosody. In this experiment, we compared the performances of all comparable algorithms by using two kinds of acoustic features: prosody and Mel-spectrogram. For each comparable methods in this and the rest experiments, their parameters were tuned individually to achieve its best performance.

The results in Table 2 show that using Mel-spectrogram as features can achieves higher accuracy than using prosody for all approaches. This proves that the spectral representation approach, especially Mel-spectrogram, is suitable for the task of speech emotion recognition. Therefore, Mel-spectrogram was selected to perform all comparisons in the next experiment. The results for KSVD-based methods are quite similar except the case of LP-KSVD with prosody feature because the number of samples in the dataset is limit. Consequently, they cannot achieve much effect from their constraints and then achieve almost the same results.

4.3 Effect of Different Sizes of Training Data

To evaluate the effects of various sizes of the training sample per class on the recognition rate and computational time, the training sample were randomly selected to 5, 10, 20, 50,
per class

can obtain the recognition result easily from the activation
discriminability of the learned dictionary. Furthermore, we
ear classifier from the training step is shared to the testing
the joint NMF. To learn dictionary jointly, the learned lin-
mantle information on the acoustic data, the locality pre-
tion, which can be achieved the advantages of both mani-
This study proposes LP-JNMF for speech emotion recogni-

5. Conclusion

This study proposes LP-JNMF for speech emotion recogni-
which can be achieved the advantages of both mani-
ment, which is proper for the large-

Acknowledgments

This research is partially supported by the Ministry of
Science and Technology under Grant Number 108-2634-F-
008-004 through Pervasive Artificial Intelligence Research
(PAIR) Labs, Taiwan.

References

[1] M.E. Ayadi, M.S. Kamel, and F. Karray, “Survey on speech emo-
tion recognition: Features, classification schemes, and databases,”
Pattern Recogn., vol.44, no.3, pp.572–587, 2011.
[2] A. Dhall, R. Goeree, J. Joshi, and T. Gedeon, “Emotion Recognition in the Wild Challenge 2016,” Proc. ICMI, pp.587–588, 2016.
[3] D. McDuff, A. Karbon, A. Kapoor, A. Roseway, and M. Czerwinski,
“AffectAura: Emotional wellbeing reflection system,” Proc. Int.
Conf. Pervasive Comput. Technol. Healthcare, pp.190–200, 2012.
[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollia,
W. Fellens, and J.G. Taylor, “Emotion Recognition in Human-
computer Interaction,” IEEE Signal Process. Mag., vol.18, no.1,
pp.32–80, 2001.
[5] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based
speech emotion recognition,” Proc. ICASSP, vol.2, pp.II–1–4, 2003.
[6] S. Lazebnik and M. Raginsky, “Supervised learning of quantizer
codebooks by information loss minimization,” IEEE Trans. Pattern
Anal. Mach. Intell., vol.31, no.7, pp.1294–1309, 2009.
[7] X.-C. Lian, Z. Li, C. Wang, B.-L. Lu, and L. Zhang, “Probabilis-
istic models for supervised dictionary learning,” Proc. CPVR,
pp.2305–2312, 2010.
[8] Q. Zhang and B. Li, “Discriminative K-SVD for dictionary learning in
face recognition,” Proc. CPVR, pp.2691–2696, 2010.
[9] T. Komatsu, Y. Senda, and R. Kondo, “Acoustic event detection
based on nonnegative matrix factorization with mixtures of local
dictionaries and activation aggregation,” Proc. ICASSP, pp.2259–2263,
2016.
[10] A. Mesaros, T. Heitola, O. Dikmen, and T. Virtanen, “Sound event
detection in real life recordings using coupled matrix factorization
of spectral representations and class activity annotations,” Proc.
ICASSP, pp.151–155, 2015.
[11] D. Kim, S.-Y. Lee, and S.-I. Amari, “Representative and discrim-
inant feature extraction based on NMF for emotion recognition in
speech,” Lecture Notes in Computer Science, vol.5863 LNCS PART
1, Heidelberg: Springer Berlin Heidelberg, pp.649–656, 2009.
[12] P. Song, S. Ou, W. Zheng, Y. Jin, and L. Zhao, “Speech emotion
recognition using transfer non-negative matrix factorization,” Proc.
ICASSP, pp.5180–5184, 2016.
[13] Z. Wu, E. Chng, and H. Li, “Joint nonnegative matrix factoriza-
tion for exemplar-based voice conversion,” Multimed. Tools Appl.,
vol.74, no.22, pp.9943–9958, 2015.
[14] S.-W. Fu, P.-C. Li, Y.-H. Lai, C.-C. Yang, L.-C. Hsieh, and Y.
Tsao, “Joint dictionary learning-based nonnegative matrix factoriza-
tion for voice conversion to improve speech intelligibility after oral
surgery,” IEEE Trans. Biomed. Eng., vol.64, no.11, pp.2584–2594,
2017.
[15] Y.K. Yiáslmaz, A.T. Cemgil, and U Şimşekli, “Generalised coupled
tensor factorisation,” Proc. NIPS, pp.2151–2159, 2011.

Table 3 Average recognition rates (%) and their standard deviations of
the comparable methods performing on different sizes of training sample
per class

| Method   | Number of training samples per class |
|----------|--------------------------------------|
|          | 5       | 10      | 20      | 50      | 100     | 200     | 480     |
| D-KSVD   | 49.71   | 45.38   | 55.26   | 78.86   | 85.60   | 87.62   | 88.15   |
|          | (6.29)  | (7.60)  | (6.97)  | (3.72)  | (2.34)  | (1.86)  | (3.48)  |
| LC-KSVD  | 49.71   | 45.38   | 55.26   | 78.86   | 85.60   | 87.58   | 88.16   |
|          | (6.29)  | (7.60)  | (6.97)  | (3.72)  | (2.34)  | (1.87)  | (3.00)  |
| LP-KSVD  | 49.71   | 45.38   | 55.26   | 78.86   | 85.60   | 87.63   | 88.23   |
|          | (6.29)  | (7.60)  | (6.97)  | (3.72)  | (2.34)  | (1.91)  | (3.39)  |
| J-NMF    | 65.05   | 75.36   | 81.70   | 84.79   | 87.59   | 87.62   | 88.47   |
|          | (4.63)  | (1.72)  | (0.82)  | (2.11)  | (1.56)  | (1.86)  | (2.15)  |
| Gr-NMF   | 68.13   | 67.34   | 69.48   | 73.30   | 77.30   | 77.75   | 77.87   |
|          | (8.56)  | (7.54)  | (7.22)  | (5.29)  | (5.27)  | (2.45)  | (3.38)  |
| LP-JNMF  | 68.49   | 76.34   | 82.38   | 87.23   | 88.76   | 88.92   | 90.17   |
|          | (4.53)  | (1.33)  | (1.48)  | (1.79)  | (1.08)  | (1.29)  | (1.55)  |

Table 4 Average computational times (ms) of the comparable methods
performing on different sizes of training samples per class

| Method   | Number of training samples per class |
|----------|--------------------------------------|
|          | 5       | 10      | 20      | 50      | 100     | 200     | 480     |
| D-KSVD   | 1.65    | 4.03    | 4.39    | 8.68    | 11.74   | 13.44   | 16.63   |
|          | (0.71)  | (0.42)  | (0.47)  | (0.88)  | (1.23)  | (1.50)  | (1.93)  |
| LC-KSVD  | 1.74    | 4.13    | 4.45    | 7.26    | 12.32   | 14.06   | 17.13   |
|          | (0.75)  | (0.42)  | (0.47)  | (0.88)  | (1.23)  | (1.50)  | (1.93)  |
| LP-KSVD  | 1.56    | 3.94    | 4.26    | 7.23    | 11.87   | 13.73   | 16.94   |
|          | (0.74)  | (0.43)  | (0.46)  | (0.87)  | (1.22)  | (1.49)  | (1.89)  |
| J-NMF    | 20.22   | 20.81   | 20.97   | 20.70   | 18.65   | 20.02   | 20.37   |
|          | (2.33)  | (2.46)  | (2.50)  | (2.48)  | (2.44)  | (2.46)  | (2.48)  |
| Gr-NMF   | 6.48    | 7.68    | 7.56    | 8.68    | 7.60    | 5.16    | 5.32    |
|          | (1.33)  | (1.48)  | (1.79)  | (1.08)  | (1.29)  | (1.55)  | (1.55)  |
| LP-JNMF  | 20.30   | 21.05   | 21.70   | 20.12   | 18.70   | 21.28   | 21.48   |
|          | (1.08)  | (1.33)  | (1.48)  | (1.79)  | (1.08)  | (1.29)  | (1.55)  |
[16] X. He and P. Niyogi, “Locality preserving projections,” Proc. NIPS, pp.153–160, 2003.

[17] J.-C. Lin, C.-H. Wu, and W.-L. Wei, “Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition,” IEEE Trans. Multimedia, vol.14, no.1, pp.142–156, 2012.

[18] Z. Jiang, Z. Lin, and L.S. Davis, “Label consistent K-SVD: Learning a discriminative dictionary for recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.11, pp.2651–2664, 2013.

[19] Y.-S. Lee, C.-Y. Wang, S. Mathulaprangsan, J.-H. Zhao, and J.-C. Wang, "Locality-preserving K-SVD based joint dictionary and classifier learning for object recognition," Proc. ACM’MM, pp.481–485, 2016.

[20] D. Cai, X. He, J. Han, and T.S. Huang, “Graph regularized non-negative matrix factorization for data representation.” IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.8, pp.1548–1560, 2011.