ACOUSTIC FEATURE EXTRACTION BY TENSOR-BASED SPARSE REPRESENTATION FOR SOUND EFFECTS CLASSIFICATION

Xueyuan Zhang, Qianhua He, Xiaohui Feng

School of Electronic and Information Engineering,
South China University of Technology, Guangzhou, Guangdong, China, 510640
eexhfeng@scut.edu.cn

ABSTRACT

This paper describes a method to extract time-frequency (TF) audio features by tensor-based sparse approximation for sound effects classification. In the proposed method, the observed data is encoded as a higher-order tensor and discriminative features are extracted in spectrotemporal domain. Firstly, audio signals are represented by a joint time-frequency-duration tensor based on sparse approximation; then tensor factorization is applied to calculate feature vectors. The three arrays of the proposed tensor are used to represent frequency, time and duration of transient TF atoms respectively. Experimental results show that exploiting tensor representation allows to characterize distinctive transient TF atoms, yielding an average accuracy improvement of 9.7% and 12.5% compared with matching pursuit (MP) and MFCC features.

Index Terms— sparse approximation, tensor factorization, sound classification, time-frequency features

1. INTRODUCTION

This paper addresses the feature extraction from tensor-based sparse representation of sounds. In the context of time-frequency (TF) feature extraction, sparse approximation techniques have become standard tools. They allow to decompose the observed signal into a small number of elementary TF atoms selected from an over-complete dictionary to achieve reconstruction with minimal distortion. The TF atoms are either dilated and translated versions of a mother function, providing joint TF localization, or learnt from training data with clustering methods such as K-means singular value decomposition (K-SVD). Given the dictionary, many algorithms have been proposed for finding the sparse coefficients for approximation, such as matching pursuit (MP), orthogonal matching pursuit (OMP), gradient pursuit (GP), basis pursuit (BP) and least absolute shrinkage and selection operator (LASSO) [1]. Despite the fact that much work has been done for dictionary training and calculating sparse coefficients, no work has ever been devoted into deriving representing features from the sparse coefficients for classification

In this work, to explore the coefficients, a tensor-based sparse representation is proposed to preserve distinctiveness of TF atoms. The three arrays in tensor respectively represent centre frequency, temporal centre position and duration of TF components. Thus the tensor is a joint time-frequency-duration expression of non-stationary signal. Discriminative features are derived from tensor factorization in spectrotemporal domain.

2. RELATION TO PRIOR WORK

The general idea is representing sparse approximation in tensor structure and factorizing this tensor to produce audio features. Related sparse approximation and tensor-based representation techniques are reviewed in this section.

Several work has been proposed to utilize the sparse approximation for sound classification. Zubair and Wang [2] directly utilize the sparse coefficients as features to train the model for signal classification. In their future work [1], max- and average- pooling operations are adopted to produce more robust feature. Chu et al [3] calculated the mean and standard deviation of frequency and scale parameters of selected atoms to produce the 4-dimensional feature. Following this idea, Sivasankaran and Prabhu [4] utilized coefficients of atoms as weights to calculate the weighted mean and standard deviation of parameters. The statistics describes the general distribution of atom parameters but distinctiveness of atoms is lost. In the proposed method, such distinctiveness is preserved in a tensor structure. Thus there is no compression in coefficients of atoms.

Representing inherent structure of the signal by tensor has been applied in various scenarios. Wu and Zhang [5] regard cochlear spectrograms of different speakers as tensors and subspaces are decomposed from the tensor to

This work was supported by the Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (2012LYM_0012), the Fundamental Research Funds for the Central Universities, South China University of Technology, China (2013ZM0090), the National Natural Science Foundation of China (61401161, 61301300)

encode the speech. Encoding coefficients are used as features for speaker classification. Sivalingam et al [6] encodes image on region descriptor matrices in tensor format to keep the descriptors in their original space. Zhang and Jiang [7] improved their scheme by presenting a discriminative dictionary learning algorithm. In the cited approaches, the dictionary atoms are derived from training samples, and tensor contains the weight vectors for different identities or regions. However, no research has been done to explore relations among the atoms. In the proposed tensor, each array represents a parameter of atoms, resulting in a tensor revealing contribution of individual parameters of atoms and a co-occurrence relation among them.

3. PROPOSED FEATURE

The proposed feature extraction framework involves following steps: a tensor is constructed to represent parameters and coefficients of sparse approximation; the tensor is decomposed to produce feature vectors. The process is shown in Fig 1.

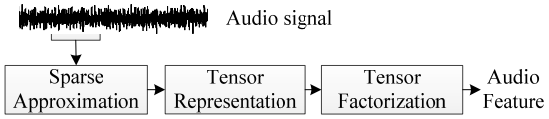


Fig 1. Flowchart of the proposed feature extraction scheme

3.1. Sparse approximation

Gabor atoms are adopted in this work which is defined as

$$g_{w,\mu,\sigma}(m)=\frac{\lambda_{w,\mu,\sigma}}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(m-\mu)^2}{2\sigma^2}\right\}\cos\left[2\pi w(m-\mu)\right] \quad (1)$$

where w, μ and σ represent central frequency, temporal position and duration respectively, $w\in\{w_1,...,w_i,...,w_I\}$, $\mu\in\{\mu_1,...,\mu_j,...,\mu_J\}$ and $\sigma\in\{\sigma_1,...,\sigma_k,...,\sigma_K\}$. m is used for denoting time samples. λ normalizes the atom to unit energy. An atom of the dictionary is implemented with parameters chosen from the parameter sets. The total number of combination is IJK, which defines the size of dictionary N. The dictionary is,

$$\begin{aligned}\boldsymbol{D}&=\left[g^{(1)},g^{(2)},...,g^{(n)},...,g^{(N)}\right]\\&=\left[g_{w_1,\mu_1,\sigma_1},g_{w_1,\mu_1,\sigma_2},...,g_{w_i,\mu_j,\sigma_k},...,g_{w_I,\mu_J,\sigma_K}\right]\end{aligned} \quad (2)$$

where $n=(i-1)JK+(j-1)K+k$ in $g^{(n)}$ is the index of Gabor atom in dictionary, and i, j and k in g_{w_i,μ_j,σ_k} are indices of parameters. The sparse approximation is represented as,

$$\begin{aligned}\boldsymbol{z}&=\boldsymbol{D}\boldsymbol{x}+\boldsymbol{e}\\&=\sum_{n=1}^{N}x_n g^{(n)}+\boldsymbol{e}\\&=\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{I}x_n g_{w_i,\mu_j,\sigma_k}+\boldsymbol{e}\end{aligned} \quad (3)$$

where $\boldsymbol{z}$ is the observed data, $\boldsymbol{e}$ is the residual, $\boldsymbol{x}$ is the sparse coefficient vector, and x_n is the n-th element in $\boldsymbol{x}$.

3.2. tensor-based sparse representation

A 3-dimensional tensor $\underline{\mathbf{T}}\in\mathbb{R}^{I\times J\times K}\geq 0$ containing IJK elements is built, in which the indices of one tensor element $t_{i,j,k}$ are (i,j,k). The indices build up a one-to-one mapping between tensor element and Gabor atom. For example, there exist a mapping between tensor element $t_{i,j,k}$ and Gabor atom g_{w_i,μ_j,σ_k} because both of their indices are (i,j,k). This mapping rearrange Gabor atoms to the higher-order tensor format. The three arrays in tensor respectively represent the indices of w, μ and σ in Gabor atoms, as shown in Fig 2.

As shown in (3), each Gabor atom g_{w_i,μ_j,σ_k} is associated with a coefficient x_n which indicates the contribution of the Gabor atom to reconstruction. Its absolute value $|x_n|$ indicate the intensity of the TF component within signal. The absolute value of x_n is then assigned to tensor element $t_{i,j,k}$. Then the tensor element position represent indices of Gabor function parameters while tensor element values represent their strength in observed signal. The whole process is shown in Fig 2.

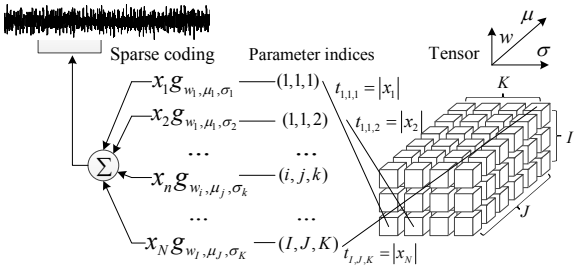


Fig 2. Tensor-based sparse representation

As shown in Fig 2, atoms and their associated weights in sparse approximation are mapped to the tensor. The one-to-one mapping is represented by line segments and the atoms and tensor elements are connected through their indices. The tensor size is depicted as fixed for illustration. The produced tensor representation has following properties:

1, tensor preserves distinctiveness of all selected atoms in sparse representation. Each dictionary atom associated with a nonzero coefficient results in a nonzero tensor element. Thus all selected atoms are represented by nonzero tensor elements, as shown in Fig 3 (c) and (d).

2, tensor integrates different parameters as different dimensions and expresses a co-occurrence relation among them. For example, the tensor element $t_{i,j,k}$ with value $|x_n|$ suggests that at time μ_j, a component with frequency w_i in length σ_k has a strength of $|x_n|$ in observed signal. The occurrence of frequency values are bonded with duration values so that each tensor element expresses a co-occurrence of them. Accordingly, the tensor describes a joint time-frequency-duration distribution.

3.3. tensor factorization

Parallel factor analysis (PARAFAC), also known as Canonical decomposition (CANDECOMP) or simply CP, decompose a M-mode tensor into the summation of a pre-specified number of outer product of M vectors. However the tensor in this work is sparse as a result of sparse approximation. And PARAFAC performs poorly when the target tensor is sparse [8]. Thus a l_1-norm regularized non-negative tensor factorization proposed by Liu et al [9] is adopted for tensor factorization. Still, because the number of nonzero elements in tensor is very small, to yield a simple representation of the underlying structure, the factorization is in the form of

$$\underline{\mathbf{T}} \approx \boldsymbol{c}_w \otimes \boldsymbol{c}_\mu \otimes \boldsymbol{c}_\sigma \tag{4}$$

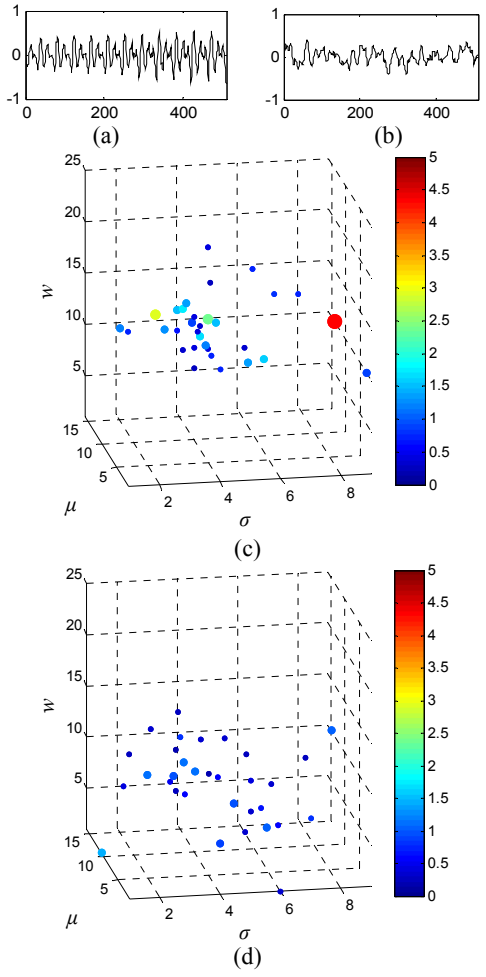
where symbol $\otimes$ denotes outer product [9] and vectors $\boldsymbol{c}_w$, $\boldsymbol{c}_\mu$ and $\boldsymbol{c}_\sigma$ represent frequency, time and duration component respectively. Vector $\boldsymbol{c}_w$ characterizes the salient frequency values within data while $\boldsymbol{c}_\sigma$ shows their durations. $\boldsymbol{c}_\mu$ indicates the activation of them but it is affected by framing and on-set shift. Thus only $\boldsymbol{c}_w$ and $\boldsymbol{c}_\sigma$ are concatenated to produce the final audio features.

4. EXPERIMENTS AND RESULTS

4.1 proposed features

The parameters are set as $w \in \{0.5\times(m/25)^{2.6}:1\le m\le 25, m\in \mathbb{Z}\}$, $\mu \in \{1+32m, 0\le m\le 15, m\in\mathbb{Z}\}$ and $\sigma \in \{2^m:3\le m\le 11, m\in \mathbb{Z}\}$. Therefore the dimensions of tensor $\underline{\mathbf{T}}$ and produced features are 25×16×9 and 34 respectively. In sparse approximation, OMP is adopted where different densities are specified. Density is defined as the number of selected atoms for sparse approximation.

A frame of female speech and river sound is illustrated in Fig 3 (a) and (b). Their tensor representations are shown in (c) and (d) respectively where the density is 32. Dots in bigger size or warmer color represent larger tensor element values. On the contrary, smaller dots or colder color represent smaller tensor element values. Tensor elements in zero values are not shown. As can be seen from (c), a notable big red dot is at location (10,14,9) of (w,μ,σ). This suggests a frequency component of 738 Hz with nearly uniform energy envelop and the temporal peak is around 417-th sample point within the short-term frame, which is consistent with the signal shown in (a). Other dots in (c) has much lower energy compared with the highest energy dot. On the other hand, dots in (d) tend to have similar and small energy values, and they are more spread out. This is consistent with the fact that river sound is more noise like than speech. Fig (e), (f) and (g) show the female speech factors $\boldsymbol{c}_w$ $\boldsymbol{c}_\mu$ and $\boldsymbol{c}_\sigma$ from factorizing the tensor in (c). Fig (h), (i) and (j) show the river sound factors $\boldsymbol{c}_w$ $\boldsymbol{c}_\mu$ and $\boldsymbol{c}_\sigma$ from factorizing the tensor in (d). As can be seen from (e) and (h), the frequency factor of female speech is more centralized than river, although they share a same frequency peak. Fig (g) and (j) show that the duration of components in female speech is much longer than that in river sounds, which is consistent with the fact that speech is periodic while river sound is non-periodic. This can also be observed from Fig (f) and (i) where the activation of speech is more centralized than river sound.



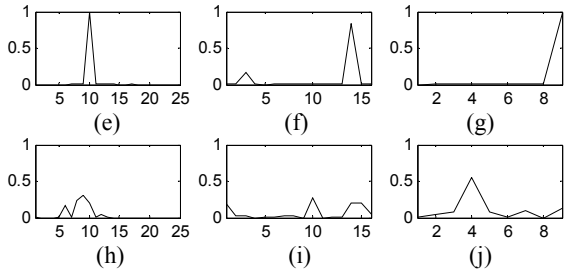


Fig 3. A frame of female speech recording (a) and river sound recording (b), their tensor representations in (c) and (d) and their factorized components in (e) to (j).

4.2 Sound effects classification

The sound effects data set contains 13 sound types including male, female, footsteps, applause, cats, birds, thunder, rivers, gunshots, engines, alarms, pianos and drums. Samples are collected from Digital Juice Sound FX Library I and II [10] and BBC Sound Effects Library [11], segmented into 3s to 10s in length, mono and sampled at 16k Hz. 8ms shifting windows of length 32ms were applied. Gaussian mixture model (GMM) with 8 mixtures are used as classifiers. All experiments are conducted in Matlab version 7.14.0 and all tests were performed on two Intel Core i7 CPUs both at 3.4GHz and 8GB RAM computer.

The proposed tensor-based feature (Tensor) is compared with MP-based feature (MP), conventional 39-dimensional MFCC (0th-order to 12th-order coefficients, delta coefficients and delta-delta coefficients), combined MP and MFCC (MP+MFCC). Classification accuracy is the ratio of correctly classified samples to total number of samples, averaged from the 10-cross validation. The average accuracy is shown in Fig 4.

As the baseline performance, the best classification accuracy that MP gives is 60.9% at the density of 4. This is consistent with the results obtained by Chu [3] where density of 5 showed best results. The best performance of MP+MFCC is 65.8% at density of 8. The proposed feature Tensor does not show improvement at small densities, but outperforms MP, MFCC and MP+MFCC when the density is larger than or equal to 16. This implies that Tensor can discover distinguishing information among multiple selected atoms. The reason for poor performance of Tensor at low density is probably the tensor is too sparse thus over-fitting occurs. The best classification accuracy 73.3% is shown by Tensor+MFCC at density of 32. Note that the densities are on a base-2 logarithmic scale.

The classification accuracy of each category is shown in Table 1, where the optimal density of each feature is chosen according to Fig 4. Tensor feature alone gives the best performance over MP, MP+MFCC and MFCC features on female, cats ,thunder, rivers, engines, gunshots, pianos and drums. By combining MFCC, classification accuracies of male, birds, footsteps, applause, rivers gunshots, alarms, pianos and drums are further improved.

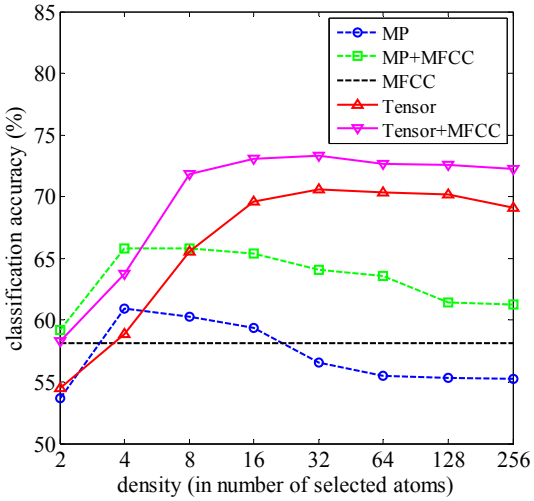


Fig 4. Classification accuracy with different densities

Table 1. Classification accuracy of each category (%)

Category	MP	MP +MFCC	MFCC	Tensor	Tensor +MFCC
male	81.4	88.6	84.7	85.4	90.7
female	77.2	84.0	83.8	87.0	91.8
cats	66.3	81.4	78.1	89.1	87.6
birds	65.7	81.3	77.7	79.3	84.0
footsteps	67.6	76.7	61.5	71.1	77.6
applause	67.5	74.9	53.9	69.9	77.0
thunder	33.9	24.5	24.4	51.3	51.3
rivers	66.8	66.1	53.5	66.8	67.7
engines	42.7	43.2	37.0	58.5	56.3
gunshots	36.5	38.7	31.5	59.7	62.4
alarms	69.4	72.2	63.3	70.4	73.6
pianos	69.5	71.3	65.0	74.3	76.1
drums	47.6	52.9	40.4	55.0	56.7
average	60.9	65.8	58.1	70.6	73.3

5. CONCLUSIONS

A novel audio time-frequency feature extraction scheme that represents transient components with tensor is proposed for sound effects classification. Since this feature utilizes the high dimensionality of tensor to represent different parameters of atoms, frequency, temporal and duration properties of transient components are preserved in the tensor structure. Representative time-frequency features are derived from factorizing the tensor. In the experiments, the proposed feature outperforms traditional MFCC and MP features.

6. REFERENCES

[1] S Zubair, F Yan, and W Wang, "Dictionary learning based sparse coefficients for audio classification with max and average pooling," *Digital Signal Processing*, vol. 23, no.3, pp. 960-970, 2013.

[2] S Zubair and W. Wang,"Audio classification based on sparse coefficients," *Sensor Signal Processing for Defence (SSPD 2011)*, London, UK, Sep. 27-29, 2011, IET, pp. 1-5

[3] S Chu, S Narayanan, and C.-C.J. Kuo, "Environmental Sound Recognition With Time–Frequency Audio Features," *IEEE Trans. Audio, Speech, and Language Process*, vol. 17 no. 6, pp.1142-1158, 2009.

[4] S. Sivasankaran,K.M.M.Prabhu,"Robust features for environmental sound classification", *2013 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 7-19 Jan. 2013, IEEE, pp. 1-6.

[5] Qiang Wu and Liqing Zhang, "Auditory sparse representation for robust speaker recognition based on tensor structure," *EURASIP Journal on Audio, Speech, and Music Processing* 2008.

[6] R. Sivalingam, D. Boley, V. Morellas et al, "Tensor sparse coding for region covariances," *In Computer Vision–ECCV 2010*, Heraklion, Crete, Greece, 5-11 Sep., 2010, Springer, pp. 722-735

[7] Yangmuzi Zhang, Zhuolin Jiang, and Larry S. Davis, "Discriminative Tensor Sparse Coding for Image Classification," *In Proceedings British Machine Vision Conference*, Bristol UK, 9-13 Sept 2013, BMVA Press, pp. 83.1-83.11

[8] K Takeuchi, K Ishiguro, A Kimura, et al, "Non-negative multiple matrix factorization," *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, Beijing, China, 3-9 Aug, 2013, AAAI Press, pp: 1713-1720.

[9] Ji Liu, Jun Liu, Peter Wonka and Jieping Ye, "Sparse non-negative tensor factorization using columnwise coordinate descent," *Pattern Recognition*, Vol 45, Issure 1, pp. 649-656, 2012

[10] Digital Juice, Inc., "The Digital Juice Sound FX Library" http://www.digitaljuice.com, accessed March 2009.

[11] British Broadcasting Corporation (BBC), "BBC Sound Effects Library," http://www.sound-ideas.com/bbc.html, accessed May 2010.