Application of Collaborative Representation in Audio Signal Recognition

Qun. Feng. Huang* and Xiao. Jiang. Li
Zhonghuan Information College, Tianjin University of Technology, Tianjin, China
*Corresponding author Email: tg778899@xzcstudio.com

ABSTRACT: Feature extraction and classifier design are the main problems of acoustic signal recognition algorithms. In this paper, we extract the mel-frequency cepstral coefficients of the acoustic features of vehicles in complex scenes. Collaborative representation is introduced for the design of a classification scheme (Collaborative Representative Classification, CRC), which synthetically considers the relationship among samples. Experiments show that the proposed algorithm produces good performance in vehicle recognition for the case of a complex data set. Compared with other classification algorithms, the method improves the precision of recognition.

1. INTRODUCTION
Recently, sparse coding or sparse representation[1] (Zhang et al., 2011) has been widely studied to solve signal recognition, partially due to the progress of $l_0$-norm and $l_1$-norm minimization techniques. In pattern classification[2], sparse representation can also be used to classify a signal and detect a target.

Sparse representation codes a signal $y$ over a dictionary $\phi$, such that $y \approx \phi \alpha$ where $\alpha$ is a sparse vector. The sparsity of $\alpha$ can be measured by $l_0$-norm, which counts the number of non-zeros in $\alpha$. Because the combinatorial $l_0$-norm is NP-hard, the $l_1$-minimization, as the closest convey function to $l_0$-norm minimization, is widely employed in sparse coding: $\min_\alpha \left\| \alpha \right\|_1 \text{ s.t. } \left\| y - \phi \alpha \right\|_2 \leq \varepsilon$, where $\varepsilon$ is a small constant. Although $l_1$-minimization is much more efficient than $l_0$-minimization, it is still time-consuming, and hence many fast algorithms have been proposed to speed up the $l_1$-minimization process.

Most literature[3]-[4] (e.g. Kopparapu & Laxminarayana, 2010; Boucheron et al., 2012) places too much emphasis on the role of $l_1$-norm sparsity in signal classification, while the role of collaborative representation, using the training samples from all classes to represent the query sample $y$, is largely ignored.

In this paper we propose a new classification scheme, namely Collaborative Representation Classification (CRC)[5], which has significantly less complexity than Sparse Representation Classification (SRC)[6], but leads to very competitive classification results. Section 2 briefly reviews the Mel-Frequency Cepstral Coefficient (MFCC) algorithms (Gang et al., 2010) that are used to extract the audio features. Section 3 briefly reviews SRC (Yang et al., 2011; Kua et al., 2011) and Section 4 presents the CRC scheme (McLaughlin et al., 2017; Larrain et al., 2017; Feng & Zhou, 2016). Section 5 presents the results of extensive experiments, and Section 6 concludes the paper.
2. MFCC ALGORITHMS
MFCC algorithms involve the following steps when used for parameter extraction:

1) Endpoint detection
   For an original audio signal, we need to detect the starting point and termination point, and then delete the silent segment. In this way, we can reduce the amount of calculation and improve the accuracy of feature extraction.

2) Pre-emphasis
   In order to obtain a flat signal spectrum, which can make spectrum analysis easier, we usually achieve it by the application of a digital pre-emphasis filter $H(z)$ after the signal has been sampled and quantized:
   \[ H(z) = 1 - a \ast z^{-1} \quad (0.9 \leq a \leq 1) \quad (1) \]

3) Windowing
   Because the audio signal is a typical time-varying signal, we use a Hamming window to cut out $N$ audio signal samples. Every sample continues for about 10 ms, so we get a series of approximately stable signals.
   \[ w(n) = 0.54 - 0.46 \cos(2\pi n/(N - 1)) \quad n = 2, 3, \ldots, N - 1 \quad (2) \]

4) Fast Fourier Transform (FFT) conversion and energy spectrum
   The audio signal is fast and unstable in the time domain, so we usually convert the signal to the frequency domain by using a FFT conversion to obtain the spectrum and energy spectrum of the signal.

5) MFCC parameter and differential cepstrum
   We use 24 triangular window filters for signal filtering, to simulate the masking effect of the human ear. We take the log of the output and then do a discrete cosine transformation. After this, we obtain the standard MFCC parameter; in order to better reflect the dynamic characteristic of the acoustic signals and improve recognition accuracy, we need the differential cepstrum:
   \[ d(n) = \frac{1}{\sqrt{\sum_{i=-k}^{k} i \times c(n + i)}} \quad (3) \]

3. THE SRC SCHEME
Natural signals, such as sound, image or seismic data, can be stored in compressed form, in terms of their projection, given a suitable basis. When the basis is chosen properly, a large number of projection coefficients are either zero or small enough to be ignored. If a signal has only $s$ non-zero coefficients, it is said to be $s$-sparse. If a large number of projection coefficients are small enough to be ignored, then the signal is said to be compressible. The signal acquisition model of Compressed Sensing (CS) [7] is quite similar to a conventional sensing framework. If $\chi$ represents the signal to be sensed, then the sensing process may be represented as:
   \[ Y = \Phi \chi \quad (4) \]
   where $\Phi$ is an $m$-by-$n$ measurement matrix and $Y$ is a measurement vector. The signal of interest $\chi$ can be expressed on a representational basis as:
   \[ \Psi x = \chi \quad (5) \]
   where $x$ is the $s$-sparse vector, representing projection coefficients of $\chi$ on $\Psi$. Measurement vector $Y$ can now be rewritten in terms of $x$ as:
   \[ Y = \theta x \quad (6) \]
where $\theta = \Phi \Psi$ is an $(m \times n)$-dimensional, reconstruction matrix. In a signal reconstruction model, we need to solve $Y = \Theta \alpha$. However, in a signal classification model, instead of recovering the signal, we need to find the appropriate category for the signal; we call this Sparse Representation Classification (SRC). The SRC algorithm involves the following steps:

1) Normalize the columns of $X$ to have unit $l_2$-norm.
2) Code $y$ over $X$ via $l_1$-minimization:

$$\hat{\alpha} = \arg \min_{\alpha} \| \alpha \|_1 \text{ s.t. } \| y - X \alpha \|_2 < \epsilon$$

(7)

where constant $\epsilon$ is to account for the dense small noise in $y$, or to balance the coding error of $y$ and the sparsity of $\alpha$.

3) Compute the residuals:

$$e_i(y) = \| y - X_i \hat{\alpha}_i \|_2$$

(8)

where $\hat{\alpha}$ is the coding coefficient vector associated with class $i$.

4) Output the identity of $y$ as:

$$\text{identity}(y) = \arg \min_{i} \{ e_i \}$$

(9)

From the above, we can see that the SRC involves two key aspects: first, the coding vector of query sample $y$ is required to be sparse; second, the coding of $y$ is performed collaboratively over the dataset $X$, instead of each subset $X_i$.

4. COLLABORATIVE REPRESENTATION-BASED CLASSIFICATION (CRC)

Sparsity\cite{8} is the key insight in compressed sensing – most signals admit a decomposition over a reduced set of signals from the same class. Unfortunately, there is no known algebraic solution to such an $l_1$-regularized least-squares formulation. A combined $l_1/l_2$ regularization tends to robustify/group the coefficients of the solution while enforcing sparsity. The representation obtained, seen as a linear decomposition over a pool of samples, has a structural meaning in that the residuals and the solution coefficients reveal the importance of each sample (or group thereof) for the new input query. This information is used in the classification (or assignment) of the query sample to the most appropriate class of samples (the one with minimum residual error or largest coefficient impact).

The basic Ordinary Least Squares (OLS) problem aims at optimizing:

$$\hat{\beta}_{OLS} = \arg \min_{\beta} \| y - X \beta \|^2$$

(10)

where $X$ is the data matrix with $(m \times n)$-dimensional samples and $\beta$ is the vector of coefficients from the representation of the query $y$. If $(X^TX)^{-1}$ exists, the algebraic solution is given by:

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^Ty$$

(11)

The Collaborative Representation with regularized least squares, here abbreviated as CR, solves:

$$\hat{\beta}_{CR} = \arg \min_{\beta} \| y - X \beta \|^2 + \lambda_{CR} \| \beta \|^2$$

(12)

where $\lambda_{CR}$ is a regulatory parameter. The algebraic solution becomes:

$$\hat{\beta}_{CR} = (X^TX + \lambda_{CR}I)^{-1}X^Ty$$

(13)

where $I$ is the $m \times m$ identity matrix.

Usually, the information used for classification is the residual corresponding to each class $c$: 

where $\hat{\beta}_c$ and $X_c$ are the coefficients and samples corresponding to class $c$ from the full representation of $y$, defined by the coefficients $\hat{\beta}_c$ and the training samples $X$. The classification decision is taken using:

$$\text{class}(y) = \arg \min_c r_c(y)$$

For a Collaborative Representation Classifier with regularized least squares (CRC) we use (as above):

$$\hat{\beta}_{CR} = Py, \quad P = (X^T X + \lambda_{CR} I)^{-1} X^T$$

and the regularized residuals are taken as:

$$r_c(y) = \frac{\Vert y - X_c \hat{\beta}_c \Vert}{\Vert \hat{\beta}_c \Vert}$$

The CRC decision is taken as above. $P$ does not depend on the query $y$ and can be precomputed. This confers a large computational advantage on CRC over SRC, which runs a query-dependent optimization. The CRC algorithm involves the following steps:

1) Normalize the columns of $X$ to have unit $l_2$-norm
2) Code $y$ over $X$ by $\hat{\rho} = Py$ where $P = (X^T X + \lambda I)^{-1} X^T y$
3) Compute the regularized residuals:

$$r_i = \frac{\Vert y - X_i \hat{\rho}_i \Vert}{\Vert \hat{\rho}_i \Vert}$$

4) Output the identity of $y$ as:

$$\text{Identity}(y) = \arg \min_i \{r_i\}.$$ 

5. EXPERIMENTAL RESULTS

5.1 Experimental data

All the experimental data in this paper is from the wireless sensor network database (the Acoustic Vehicle Classification Dataset) of the SITEX02 experiment of Project SensIT at DARPA: http://www.ecs.umass.edu/~mduarte/Software.html [accessed 7 December 2018]. The database contains two audio files of military vehicles: an Assault Amphibian Vehicle (AAV) and a Dragon Wagon (DW).

5.2 Description and experimental results

For the experiment we select the audio signal with a frequency of 4,960 Hz (see Figure 1), and then use the FFT and MFCC algorithms to extract features.
First, we use the SRC algorithms to classify the signal using either an FFT or an MFCC eigenvector, and then we make a comparison of the two eigenvectors (see Figure 2). For the experiment, we used 90 samples. After extracting the features, we randomly selected \( n \) samples as training samples, and used the remainder as testing samples. Finally, we compared the recognition rate.
### Table 1 Comparison of FFT and MFCC eigenvectors in signal classification.

| Eigenvector | Vehicle category | Training number |
|-------------|------------------|-----------------|
|             |                  | 20              |
|             |                  | 40              |
|             |                  | 60              |
| FFT         | AAV              | 75.14%          |
|             |                  | 73.60%          |
|             |                  | 75.33%          |
|             | DW               | 91.71%          |
|             |                  | 95.60%          |
|             |                  | 94.67%          |
| MFCC        | AAV              | 82.57%          |
|             |                  | 85.47%          |
|             |                  | 84.89%          |
|             | DW               | 91.14%          |
|             |                  | 92.93%          |
|             |                  | 93.11%          |

From Table 1, we can see that as the number of training samples is increased, the recognition rate of both eigenvectors increases slightly for each classification algorithm. In addition, we can see that the MFCC eigenvectors perform better than the FFT eigenvectors in classifying the AAV signal, and a little worse than the latter in classifying the DW signal. However, if we consider the overall performance, we can see that the MFCC eigenvectors outperform the FFT ones.

Thus, having extracted the MFCC eigenvectors, we used either the SRC or the CRC algorithm to classify the signal, and then made a comparison of the two. For the experiment, we used 90 samples. After extracting the features, we randomly selected \( n \) samples as training samples, and used the remainder as testing samples. Last, we compared the recognition rate and time cost of the two algorithms.

### Table 2 Comparison of SRC and CRC algorithms in signal classification (recognition rate and compute time).

| Classification algorithm | Vehicle category | Training number |
|--------------------------|------------------|-----------------|
|                          |                  | 20              |
|                          |                  | 30              |
|                          |                  | 40              |
|                          |                  | 50              |
| SRC                      | AAV              | 0.76            |
|                          |                  | 12.65 s         |
|                          | DW               | 0.96            |
|                          |                  | 0.95            |
|                          |                  | 31.07 s         |
|                          |                  | 0.82            |
|                          |                  | 0.88            |
|                          |                  | 0.92            |
|                          |                  | 54.14 s         |
|                          |                  | 0.90            |
|                          |                  | 0.90            |
|                          |                  | 72.94 s         |
| CRC                      | AAV              | 0.81            |
|                          |                  | 0.04 s          |
|                          | DW               | 0.91            |
|                          |                  | 0.80            |
|                          |                  | 0.06 s          |
|                          |                  | 0.78            |
|                          |                  | 0.90            |
|                          |                  | 0.06 s          |
|                          |                  | 0.80            |
|                          |                  | 0.85            |
|                          |                  | 0.07 s          |

From Table 2, we can see that as the number of training samples is increased, the time cost of both algorithms increases, while the overall recognition rate increases slightly. Although the CRC algorithm produces a slightly lower recognition rate than the SRC algorithm, it costs much less in time terms.

### 6. CONCLUSION

In this paper, we briefly introduce the acoustic signal recognition system and its component parts. Then we introduce a new acoustic signal recognition algorithm, CRC, and make a detailed comparison with the SRC method, comparing recognition rate and time cost. We find that the CRC algorithm has almost the same recognition rate as SRC, but at a much lower cost in time. However, we have not investigated further the impact that different signals may have on the result. In ongoing research, we may explore the method prior to signal extraction. Because SRC is becoming increasingly popular in signal processing, the advantage of CRC in costing much less time should give it broader application in the future.

### REFERENCES

[1] Boucheron, L.E., De Leon, P.L. & Sandoval, S. (2012). Low bit-rate speech coding through quantization of mel-frequency cepstral coefficients. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(2), 610–619.

[2] Feng, Q. & Zhou, Y. (2016). Kernel combined sparse representation for disease recognition. *IEEE Transactions on Multimedia, 18*(10), 1956–1968.
[3] Kopparapu, S.K. & Laxminarayana, M. (2010). Choice of mel filter bank in computing MFCC of a resampled speech. In 10th International Conference on Information Sciences, Signal Processing and their Applications (ISSPA 2010), Kuala Lumpur, Malaysia, 10–13 May 2010 (pp. 121–124). New York: IEEE. doi:10.1109/ISSPA.2010.5605491

[4] Kua, J.M.K., Ambikairajah, E., Epps, J. & Togneri, R. (2011). Speaker verification using sparse representation classification. In 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing: Proceedings (pp. 4548–4551). New York: IEEE. doi:10.1109/ICASSP.2011.5947366

[5] Larrain, T., Bernhard, J.S., Mery, D. & Bowyer, K.W. (2017). Face recognition using spare fingerprint classification algorithm. IEEE Transactions on Information Forensics and Security, 12(7), 1646–1657.

[6] McLaughlin, N., Ming, J. & Crookes, D. (2017). Largest matching areas for illumination and occlusion robust face recognition. IEEE Transactions on Cybernetics, 47(3), 796–808.

[7] Yang, M., Zhang, L., Feng, X. & Zhang, D. (2011). Fisher discrimination dictionary learning for sparse representation. In Proceedings of 2011 International Conference on Computer Vision (pp. 543–550). Washington, DC: IEEE Computer Society. doi:10.1109/ICCV.2011.6126286

[8] Zhang, L., Yang, M. & Feng, X. (2011). Sparse representation or collaborative representation: Which helps face recognition? In Proceedings of 2011 International Conference on Computer Vision (pp. 471–478). Washington, DC: IEEE Computer Society. doi:10.1109/ICCV.2011.6126277