A Study of Speech Feature Extraction Based on Manifold Learning

Cheng Hao1,a, Ma Xin2,b*, Yugong Xu3,c
1School of Information Science and Engineering Shandong University Qingdao China
2School of Information Science and Engineering Shandong University Qingdao China
3School of Information Science and Engineering Shandong University Qingdao China
acheng92hao@163.com, bmax@sdu.edu.cn, c1069202299@qq.com

ABSTRACT Manifold learning is a nonlinear data dimension reduction method. It can look for the essence of things from the observed phenomena, and find the inherent law of data. Traditional MFCC feature will lead a slower learning speed on account of it has high dimension and useless noise. Therefore, a speech feature extraction method based on manifold learning is proposed. Firstly, we use the manifold learning dimension reduction algorithm for the dimension reduction of Mel features and then for vowels classification. In order to further demonstrate the effectiveness of manifold learning feature in speech recognition, we propose a fusion speech feature extraction method and apply it to the identification of Chinese isolated words. Experiments prove that the fusion feature extraction method has achieved a better result than that of traditional MFCC feature extraction method.

1. INTRODUCTION
There are abundant feature parameters in speech signal, and different feature vectors represent different physical and acoustic meanings. The choice of characteristic parameters is of great importance to the success or failure of the speech recognition system. If a good feature parameter is selected, it will help to improve the recognition rate. Feature extraction is to extract or reduce the information in the speech signal which has nothing to do with the recognition, reduce the amount of data to be processed during the subsequent identification stage, and generate characteristic parameters that represent the speaker information carried in the speech signal. By far, the most commonly used speech characteristics are Linear Prediction Cepstrum Coefficient (LPCC) [1] and MEL Cepstrum coefficient (MFCC) [2]. All of them have achieved good recognition effect in speech recognition.

Due to physiological limits, speech organs only have a few degrees of freedom, people can only produce a limited range of sounds, which occupy a restricted area of acoustic space. Therefore, the speech signal data can be regarded as a low-dimensional manifold embedded in the high-dimensional space of all possible waveform. The purpose of feature extraction is to represent the most important essential rule of speech signal with as few data dimensions as possible.

Manifold learning is a nonlinear data dimension reduction method, which was first proposed in the famous science magazine Science in 2000 [3, 4]. In 2005, Jansen et al. discovered that voice data is data embedded in high dimensional acoustic space, and in 2006, the Fourier transform of voice signal was used to further prove the essence of the existence of manifold in voice signal [5]. Errity Andrew et al. [4] studied the speech signal manifold and the possibility of the use of manifold learning method for speech signal analysis. They believe that manifold learning can be used to explore the hidden low-
dimensional manifold structure in speech signal, and the manifold feature extracted by manifold learning algorithm is beneficial to improve the accuracy of ASR system. Wang li et al. [7] put the MFCC feature into the lower dimension feature space by using the method of manifold dimension reduction, and extracted the manifold features, and presented a research idea of using the manifold learning method to extract the characteristic parameters of speech signal.

The manifold learning algorithm can discover the geometric structure contained in the high dimensional data space, and the essential connection between sample data, and realize data dimension reduction [8, 9]. In this paper, manifold learning algorithm is applied to feature extraction, which can overcome dimension disaster, obtain essential features, save storage space, and eliminate useless noise.

2. MANIFOLD LEARNING ALGORITHM

Isometric Feature Mapping (ISOMAP) is a nonlinear data dimension reduction algorithm proposed by Tenenbau et al. [3] based on MDS algorithm. Assuming that the data is distributed on a manifold space, in order to keep inherent geometry relationship in high-dimensional space, the relative distance between each sample point in a high dimensional sample set usually needs to remain constant. In Isomap algorithm, the author introduced Geodesic Distance to replace the traditional Euclidean Distance, which can better maintain high dimensional sample space geometric relationship between each sample, and then obtain the optimal structure of dimension reduction. Tenenbau gives a method to approximate the geodesic distance by using the shortest path between two points in the data neighborhood graph in [3]. Bernstein et al. [8, 9] have proved that the geodetic distance can be approximated by the shortest distance when there are enough data samples and under uniform sampling conditions.

Classic ISOMAP algorithm has three steps:

1. Establish neighborhood relationship diagram $G(V, E)$: For each $x_i (i = 1, \cdots, N)$, according to certain criteria to determine K phase velocity point and calculate its K neighbor. To point $x_i$ for vertex. Euclidean distance $d_e(x_i, x_j) \in G(V, E)$.

2. According to the graph set good path to calculate the shortest path between any two points use them as myopia geodesic distance $d_g(x_i, x_j)$ to construct the geodesic distance matrix $D[d_g(x_i, x_j)]_{N \times N}$.

3. For geodesic distance matrix $D[d_g(x_i, x_j)]_{N \times N}$ use traditional MDS algorithm, looking for a low dimensional data $Y = (y_1, y_2 \cdots y_n)$.

Laplacian Eigenmaps (LE) [8, 12] is a nonlinear dimension-reduction algorithm for semi-supervised learning, which is also an algorithm to maintain the local relationship between samples. LE algorithm base on spectral graph theory, through the calculation of adjacency matrix of sample points in higher dimensional sample space and feature decomposition of the laplacian matrix of the adjacency matrix, and then calculate the nonlinear dimension reduction result of the high-dimensional sample data.

Locality Preserving Projection (LPP) [9] was proposed by He Xiaofei and Partha Niyogi in 2003. LPP algorithm is a linear approximation of LE [8, 12] algorithm. LPP algorithm calculates the transformation matrix by constructing the adjacency graph of the data set and using the Laplace matrix of the adjacency graph. By using this transformation matrix, data points can be mapped to the low-dimensional subspace.

3. FEATURE EXTRACTION OF VOICE SIGNAL

3.1 MFCC Feature

MFCC [2] feature is a kind of speech characteristic parameter based on human auditory mechanism and is an important characteristic parameter in the field of speech signal processing. According to
human auditory mechanism, the division of subjective perceptual frequency domain is not linear. Long-term studies have shown that the relationship between perceived frequency and actual frequency is

$$F_{\text{mel}}(f) = 1125 \times \ln(1 + \frac{f}{700}),$$

(1)

Where, $f$ is the actual frequency, and the unit is Hz, is the perceived frequency, and the unit is Mel. By analyzing the speech signal in the perceptual frequency domain, we can get a better approach to the auditory processing process and get a better result. According to the critical band theory, a series of frequency groups are divided on the frequency to form several triangular bandpass filter groups, namely Mel filter groups. The extraction and calculating process of MFCC features as show in Figure 1.

![Figure 1. MFCC feature extraction process](image)

### 3.2 Manifold Learning Feature

In most applications, the number of triangular filters in the Mel filter group is generally 26, so the dimension $D$ of the initial sample parameter vector of speech signal is 26. However, relevant studies have shown that the degree of freedom of speech signal is generally 5 to 6, and the dimension of initial sample parameters can be reduced by using manifold learning method, so as to obtain the eigenfeature parameters of speech signal. Therefore, a new method of extracting speech feature parameters is proposed in this paper. The extraction process of phonetic characteristic parameters based on manifold learning as shown in Figure 2.

![Figure 2. Manifold learning feature extraction process](image)

### 3.3 Feature Fusion

On the basis of vowel classification, in order to further prove the viewpoint proposed in this paper in the identification of Chinese numeral isolated words, the manifold learning features and the MFCC features are fused as new features for speech recognition. In this way, we can not only use the advantages of MFCC features based on human auditory mechanism, but also use the advantages of manifold learning features. The feature fusion method adopts series fusion, namely, the MFCC feature matrix is directly spliced behind the manifold learning feature matrix. The extraction process of fusion features is shown in Figure 3.
Training data

Testing data

Preprocessing

FFT

Spectral energy

Mel filtering energy

DCT

MFCC feature

Manifold learning algorithm

Manifold learning feature

Fusion feature

Figure 3. Fusion feature extraction process

4. EXPERIMENTS

4.1 Vowel Separability Analysis

According to the vowel phoneme map given by IPA, 5 vowels, such as aa, ae, uw, iy and eh, are selected, and 400 samples are extracted from the TIMIT [12] corpus for each vowel, and then we calculate the MFCC features and manifold learning features of the 2000 samples respectively. The manifold learning algorithm used in this paper is ISOMAP, LE and LPP, and the parameter k (the number of nearest neighbor samples) used by the algorithm is 12 by default. In order to investigate the performance of each algorithm in different dimensions, the 2-15 dimensional features of these vowel samples are calculated. FIG. 4 shows the distribution of 2-dimensional manifold feature sample points of the four vowels aa, ae, uw and iy. It should be noted that because ae and eh sound alike, the separation degree of each algorithm is not very good, so only four vowels are used for visualization. It can be seen from the figure that the manifold learning algorithm is effective in the visualization of vowel clustering analysis, while the traditional MFCC feature is not effective.

Figure 4. Distribution of manifold features in two-dimensional space
The performance of each algorithm can be measured by calculating the vowel separability in low dimensional space. In this paper, Bhattacharyya distance \[^{1,14}\] is used to measure the separability of two kinds of samples. The mean value of the two samples is \(M_1\) and \(M_2\), and the variance is \(\Sigma_1\) and \(\Sigma_2\). The calculation method of Bhattacharyya distance is as follows

\[
D_{\text{Bhattacharyya}} = \frac{1}{8}(M_2 - M_1)^T \left[ \frac{\Sigma_1 + \Sigma_2}{2} \right]^{-1}(M_2 - M_1) + \frac{1}{2} \ln \left( \frac{\Sigma_1 + \Sigma_2}{2\Sigma\Sigma} \right),
\]

(2)

The Bhattacharyya distance between each of the five vowels in 2-15 dimensional feature space is calculated. By comparison, the separation degree of the vowels in the manifold feature space is significantly better than that of the MFCC feature space. FIG. 5 shows the results in the 3-dimensional feature space, which is obviously better than the results in [6].

![Figure 5. Vowel separability test](image)

4.2 Vowel Classification Analysis

In order to test the usability of manifold learning algorithm, this section conducts a classification analysis of these five vowels. In order to investigate the performance of each algorithm in different dimensions, the 2-15 dimensional features of these five vowel samples are calculated. In the classification analysis, 300 samples of each vowel are taken and marked to form the training set, and the remaining 100 samples were used for classification test. KNN (k-nearest Neighbour) \[^{15}\] classifier was used for the classification test, and the parameter \(k=5\). FIG. 6 shows the results of kNN classification experiment. It can be seen from the figure that the classification results obtained by using the manifold features are obviously better than the traditional MFCC features. The features based on manifold learning can achieve better classification result at low dimension, and can reflect the manifold structure of speech signal. With the increase of dimension, the error rate does not decrease significantly, which means that LE and LPP algorithm cannot extract more information from local linear space.
In the above experiment, the default nearest neighbor parameter $k$ of the manifold learning dimension-reduction algorithm we used is 12, which is also one of the most important parameters of the dimension-reduction algorithm and has a great influence on the classification results of vowels. It can be seen from Figure 7 that the parameter $k$ has a significant influence on the accuracy of vowel classification. When $k=8$ or $k=16$, the accuracy is not very obvious with the change of dimension. When $k=10$, $k=12$ or $k=14$, the accuracy is increased with the increase of feature dimension. When $k=12$ and the feature dimension is 8, the accuracy of vowel classification reaches up to 92.25%.

4.3 Speech Recognition Analysis

In order to test the effect of manifold learning algorithm in continuous speech recognition, this paper uses the HMM-GMM to conduct the experiment of isolated words recognition in Chinese pronunciation. The training set uses 10 Arabic numerals from 0 to 9, 200 samples in total, 20 samples for each number, and the sampling rate is 16KHz. There are 50 samples in the test set, 5 samples for each number. In the experiment, in order to integrate the advantages of the two algorithms, the manifold learning features and the MFCC features are fused together to obtain the new fusion features for the recognition of Chinese isolated words. In this experiment, ISOMAP manifold learning algorithm is adopted, and the nearest neighbor parameter $k$ is 12. FIG. 8 is the result of the experiment, which compares the average recognition rate of fusion feature extraction method and MFCC feature extraction method in different dimensions. It can be seen from the figure that the average recognition
rate of ISOMAP fusion feature is much higher than that of traditional MFCC feature in the case of low dimension or even some high dimension. And we can see from Table 1 that the average recognition rate of MFCC and MFCC-ISOMAP are 42.7% and 52% respectively. The MFCC-ISOMAP fusion feature extraction method has better performance in Chinese isolated words recognition.

![Figure 8. 0-9 Chinese number recognition](image)

**Table 1. Average recognition rate of different feature extraction method (%)**

| Feature dimension | MFCC | MFCC-ISOMAP |
|-------------------|------|-------------|
| 2                 | 26   | 48          |
| 4                 | 28   | 42          |
| 6                 | 34   | 50          |
| 8                 | 38   | 56          |
| 10                | 46   | 54          |
| 12                | 54   | 50          |
| 14                | 50   | 64          |
| 16                | 52   | 48          |
| 18                | 56   | 56          |
| **Average recognition rate** | **42.7** | **52** |

5. CONCLUSION

In this paper, we apply the manifold learning algorithm to reduce the dimension of vowel features, and then use KNN classifier to test the manifold feature extraction method. The results show that the vowel classification error rate of manifold learning feature extraction method is much lower than that of traditional MFCC feature extraction method, which means the manifold learning feature extraction method is effective in vowel classification. In order to further demonstrate the effectiveness of manifold learning algorithm in speech recognition, the ISOMAP fusion feature extraction method is used to recognize Chinese isolated words. The HMM-GMM is used to model the speech signal. The experimental results show that the recognition rate of ISOMAP fusion feature extraction method is also higher than that of traditional MFCC feature extraction method, which indicates that the ISOMAP fusion feature extraction method is helpful to improve the recognition rate of Chinese isolated words.
Whether other manifold learning algorithms are also effective, and whether the manifold learning algorithm can be combined with CNN, RNN and other deep learning models for speech recognition is the next step of this paper. In any case, manifold learning as the latest machine learning algorithm, has obtained the certain achievement in the face recognition, which will become an important research direction to promote the breakthrough of speech recognition.

ACKNOWLEDGEMENT
This work was supported by the National Natural Science Foundation of China under Grant No.62171453. The authors would like to thank Xu Yugong for his helpful works and comments.

REFERENCES
[1] Gupta H, Gupta D. LPC and LPCC method of feature extraction in Speech Recognition System[C]/ Cloud System and Big Data Engineering. IEEE, 2016:498-502.
[2] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. Readings in Speech Recognition, 1990, 28(4):65-74.
[3] Tenenbaum J B, Silva V D, Langford J C. A Global Geometric Framework for Nonlinear Dimensionality Reduction [J]. Science, 2000, 290(5500):2319-2323.
[4] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. [J]. Science, 2000, 290(5500):2323-2326.
[5] Jansen A, Niyogi P. Intrinsic fourier analysis on the manifold of speech sounds [J]. Science China, 2006, 56(3):1-11.
[6] Errity A, Mckenna J. An investigation of manifold learning for speech analysis[C]/ INTERSPEECH 2006 - Icslp, Ninth International Conference on Spoken Language Processing, Pittsburgh, Pa, Usa, September. DBLP, 2006.
[7] Zhang P, Wang L. The application research of speech feature extraction based on the manifold learning [J]. ICCSEE-13, 2013.
[8] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering [J]. Advances in Neural Information Processing Systems, 2009, 14(6):585-591.
[9] He X. Locality preserving projections [J]. Advances in Neural Information Processing Systems, 2003, 16(1):186-197.
[10] Balasubramanian M, Schwartz E L. The isomap algorithm and topological stability. [J]. Science, 2002, 295(5552):7.
[11] Bernstein M, Silva V D, Langford J C, et al. Graph Approximations to Geodesics on Embedded Manifolds [J]. Stanford University, 2000, 24(9):153--158.
[12] Belkin M, Niyogi P. Laplacian Eigenmaps for dimensionality reduction and data representation [M]. MIT Press, 2003.
[13] Garofolo J S, Lamel L F, Fisher W M, et al. TIMIT Acoustic-phonetic Continuous Speech Corpus[C]/ Linguistic Data Consortium. 1993.
[14] Fukunaga, Keinosuke. Introduction to statistical pattern recognition (2nd ed.)[J]. 1990, 60(12-1):2133-2143.
[15] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7):2038-2048.