Manifold Learning-based Feature Transformation for Phone Classification

Andrew Errity, John McKenna and Barry Kirkpatrick

School of Computing
Dublin City University, Dublin 9, Ireland
{andrew.errity, john.mckenna, barry.kirkpatrick}@computing.dcu.ie

Abstract

This paper investigates approaches for low dimensional speech feature transformation using manifold learning. It has recently been shown that speech sounds may exist on a low dimensional manifold nonlinearly embedded in high dimensional space. A number of techniques have been developed in recent years that attempt to discover the geometric structure of the underlying low dimensional manifold. The manifold learning techniques locally linear embedding and Isomap are considered in this study. The low dimensional feature representations produced by these techniques are applied to several phone classification tasks on the TIMIT corpus. Classification accuracy is analysed and compared to conventional MFCC features and PCA, a linear dimensionality reduction method, transformed features. It is shown that features resulting from manifold learning are capable of yielding higher classification accuracy than these baseline features. The best phone classification accuracy in general is demonstrated by feature transformation with Isomap.

1. Introduction

Feature transformation is an important part of the speech recognition process and can be viewed as a two step procedure. Firstly, relevant information is extracted from short time segments of the acoustic speech signal using a procedure such as Fourier analysis, cepstral analysis or some other perceptually motivated analysis. The resulting D-dimensional parameter vectors are then transformed to a feature vector of lower dimensionality d ($d \leq D$). The aim of dimensionality reduction is to produce features which are concise low dimensional representations that retain the most discriminating information for the intended application and are thus more suitable for pattern classification. Dimensionality reduction also decreases the computational cost associated with subsequent processing.

Physiological constraints on the articulators limit the degrees of freedom of the speech production apparatus. As a result humans are only capable of producing sounds occupying a subspace of the acoustic space. Thus, speech data can be viewed as lying on or near a low dimensional manifold embedded in the original acoustic space. The underlying dimensionality of speech has been the subject of much previous research including classical dimensionality reduction analysis [1, 2], nonlinear dynamical analysis [3] and manifold learning [4]. The consensus of this work is that some speech sounds, particularly voiced speech, are inherently low dimensional.

Dimensionality reduction methods aim to discover this underlying low dimensional structure. These methods can be categorised as linear or nonlinear. Linear methods are limited to discovering the structure of data lying on or near a linear subspace of the high dimensional input space. The most widely used linear dimensionality reduction methods include principal component analysis (PCA) [5] and linear discriminant analysis (LDA) [6]. These methods have been successfully applied to feature transformation in speech processing applications [7, 8] in the past.

However if speech data occupies a low dimensional submanifold nonlinearly embedded in the original space, as proposed previously [2, 4], linear methods will fail to discover the low dimensional structure. A number of manifold learning, also referred to as nonlinear dimensionality reduction, algorithms have been developed [9–11] which overcome the limitations of linear methods. Manifold learning algorithms have recently been shown to be useful in a number of speech processing applications including low dimensional visualization of speech [4,11–14] and limited phone classification tasks [14,15].

In this paper, we build upon previous work and apply two manifold learning algorithms, locally linear embedding (LLE) [9] and isometric feature mapping (Isomap) [10], to extract features from speech data. These features are evaluated in phone classification experiments using a support vector machine (SVM) [16] classifier. The classification performance of these features is compared to baseline Mel-frequency cepstral coefficients (MFCC) and those resulting from the classical linear method, PCA.

The remainder of this paper is structured as follows. In Section 2, the manifold learning algorithms LLE and Isomap are briefly described. Section 3 details the experimental procedure, data set, parameter extraction, feature transformation and classification technique used. Results are examined and discussed in Section 4, with conclusions presented in Section 5. Finally, possibilities for future work are outlined in Section 6.

2. Manifold learning algorithms

2.1. Locally linear embedding

LLE [9] is an unsupervised learning algorithm that computes low dimensional embeddings of high dimensional data. The principle of LLE is to compute a low dimensional embedding with the property that nearby points in the high dimensional space remain nearby and similarly co-located with respect to one another in the low dimensional space. In other words, the embedding is optimised to preserve local neighbourhoods.

The LLE algorithm can be summarised in three steps:

1. For each data point X_i, compute its k nearest neighbours (based on Euclidean distance or some other appropriate definition of 'nearness').
2. Compute weights W_{ij} that best reconstruct each data point X_i from its neighbours, minimising the reconstruction error E:

$$E(W) = \sum_i \left| X_i - \sum_j W_{ij} X_j \right|^2 \qquad (1)$$

3. Compute the low dimensional embeddings Y_i, best reconstructed by the weights W_{ij}, minimising the cost function Ω:

$$\Omega(W) = \sum_i \left| Y_i - \sum_j W_{ij} Y_j \right|^2 \quad (2)$$

In step 2, the reconstruction error is minimised subject to two constraints: first, that each input is reconstructed only from its nearest neighbours, or $W_{ij} = 0$ if X_i is not a neighbour of X_j; second, that the reconstruction weights for each data point sum to one, or $\sum_j W_{ij} = 1 \; \forall i$. The optimum weights for each input can be computed efficiently by solving a constrained least squares problem.

The cost function in step 3 is also based on locally linear reconstruction errors, but here the weights W_{ij} are kept fixed while optimising the outputs Y_i. The embedding cost function in Equation (2) is a quadratic function in Y_i. The minimisation is performed subject to constraints that the outputs are centered and have unit covariance. The cost function has a unique global minimum solution for the outputs Y_i. This is the result returned by LLE as the low dimensional embedding of the high dimensional data points X_i.

2.2. Isomap

The Isomap algorithm [10] offers a differently motivated approach to manifold learning. Isomap is a nonlinear generalisation of multidimensional scaling (MDS) [6] that seeks a mapping from high dimensional space $\mathbf{X}$ to low dimensional feature space $\mathbf{Y}$ that preserves geodesic distances between pairs of data points—that is, distances on the manifold from which the data is sampled.

While Isomap and LLE have similar aims, Isomap is based on a different principle than LLE. In particular, Isomap attempts to preserve the global geometric properties of the manifold while LLE attempts to preserve the local geometric properties of the manifold.

As with LLE, the Isomap algorithm consists of three steps:

1. Construct a neighbourhood graph - Determine which points are neighbours on the manifold based on distances $l(i, j)$ between pairs of points i, j in the input space (as in step 1 of LLE). These neighbourhood relations are then represented as a weighted graph over the data points with edges of weight $l(i, j)$ between neighbouring points.
2. Compute the shortest path between all pairs of points among only those paths that connect nearest neighbours using a technique such as Dijkstra's algorithm.
3. Use classical MDS to embed the data in a d-dimensional Euclidean space so as to preserve these geodesic distances.

3. Experiments

3.1. Classification tasks

The objective of these experiments is to perform phone classification using four different feature types: baseline MFCC vectors and features produced by applying PCA, Isomap and LLE to MFCC vectors. Each feature type was evaluated in three phone classification experiments. The first experiment involves distinguishing between a set of five vowels ('aa', 'iy', 'uw', 'eh', and 'ae'). Phones are labeled using TIMIT symbols [17]. In the second test, a further five vowels ('ah', 'ay', 'oy', 'ih' and 'ow') were added to the previous vowel set, forming a more complex ten class vowel classification problem. The final test involves classifying a set of 19 phones into their associated phone classes. The phone classes and phones used were: vowels (listed above), fricatives ('s', 'sh'), stops ('p', 't', 'k'), nasals ('m', 'n') and, semivowels and glides ('l', 'y').

3.2. Data

The speech data used in this study was taken from the TIMIT corpus [17]. This corpus contains 6300 utterances, 10 spoken by each of 630 American English speakers. The speech recordings are provided at a sampling frequency of 16 kHz.

3.3. Parameter extraction

Based on the phonetic transcriptions and associated phone boundaries provided in TIMIT all units of a subset of phones, listed in Section 3.1, were extracted from the corpus. One 40 ms frame was extracted from the middle of each phone unit (units of duration less than 100 ms were discarded). The raw speech frames were amplitude normalised, preemphasized with the filter $H(z) = 1 - 0.98z^{-1}$ and Hamming windowed. Following this preprocessing, 19-dimensional MFCC vectors were computed for each frame. These MFCC vectors serve as both a baseline feature and high dimensional input for PCA, Isomap and LLE methods.

3.4. Feature transformation

For each of the three phone classification experiments, 250 units representing each of the required phones were chosen at random from those extracted above to make up the data set. PCA, Isomap and LLE were applied to the equivalent set of MFCC vectors.

In order to examine the ability of the feature transformation methods to compute concise representations of the input vectors retaining discriminating information, the dimensionality of the resulting feature vectors was varied from 1 to 19. A separate classifier was subsequently trained and tested using feature vectors with each of the 19 different dimensionalities. Thus the ability of these feature transformation methods to produce useful low dimensional features could be evaluated and changes in performance with varying dimension analysed. As a baseline the original MFCC vectors were used, also varying in dimensionality from 1 to 19.

The number of nearest neighbours, k, used in Isomap and LLE was set equal to 14 and 6 respectively. These values were chosen empirically by varying k and examining classification performance. The performance of both methods was found to be sensitive to the choice of k.

3.5. Support vector machine classification

SVM [16], a powerful classification tool, was used in these experiments. SVM is a binary pattern classification algorithm. For our experiments it is necessary to construct a multiclass classifier. This was achieved using a one-against-one training scheme, training one classifier for every possible pair of classes. The final classification result was determined by majority voting.

It is also necessary to choose an appropriate kernel function to be used in the SVM. In order to select an effective

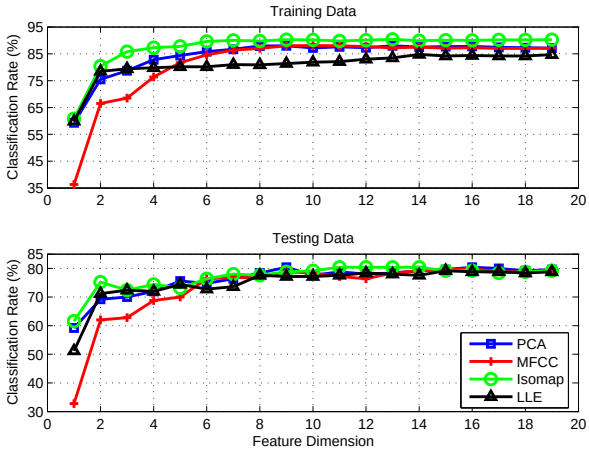

Figure 1: *Five vowel classification results for baseline MFCC, PCA, Isomap and LLE features on the TIMIT database.*

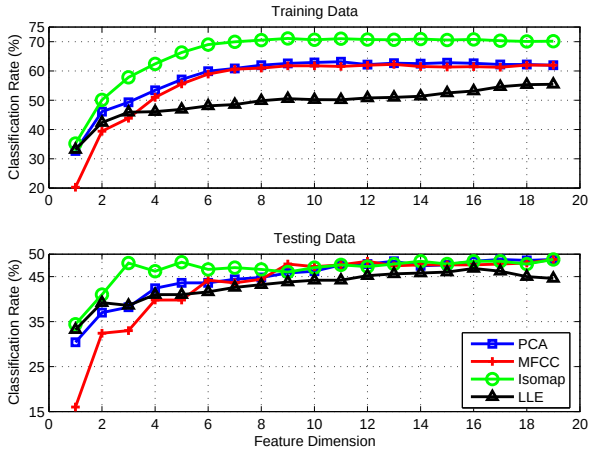

Figure 2: *Ten vowel classification results for baseline MFCC, PCA, Isomap and LLE features on the TIMIT database.*

kernel, different SVM models using linear, polynomial and radial basis function (RBF) kernels were evaluated in a number of phone classification tasks. SVM with RBF kernel demonstrated the best classification accuracy and is used for classification throughout this work. The RBF kernel used is given in Equation (3) below, with x and x' feature vectors and d the feature vector dimensionality.

$$K(x, x') = exp(-\frac{1}{d}\left\|x - x'\right\|^2) \tag{3}$$

In all classification experiments 80% of the data was assigned as training data with the remaining 20% withheld and used as unseen testing data.

4. Results

In each experiment the classifier was evaluated on each of the four feature types. The dimensionality of the feature vectors used in the experiment vary from 1 to 19—the original, full dimension. Results are presented for evaluation on both the training data and testing data.

Fig. 1 shows the results of the five vowel classification task using the baseline MFCC, PCA, Isomap and LLE features. The percentage of phones correctly classified is given on the vertical axis. The horizontal axis represents the dimensionality of the feature vector. The results in Fig. 1 can be summarized as follows:

- The performance of the baseline MFCC vectors improves with increasing dimensionality, plateauing at a dimensionality of approximately 8.
- PCA features offer improvements over baseline MFCC for low dimensions, 1 to 7.
- For the training data, maximum classification accuracy in all dimensions is demonstrated with Isomap features, outperforming all other features including the original full 19-dimensional MFCC vectors.
- Isomap features also offer performance comparable to, and in some dimensions better than, other features on the testing data.
- Accuracy with LLE features is better than both MFCC and PCA in low dimensions, ($d < 3$). However in higher dimensions LLE features do not offer a performance increase over other methods.

Results for ten vowel classification are given in Fig. 2. The results are similar to those of the task above, with reduced classification accuracy due to increased complexity and increased possibility of phone confusion. The important findings are as follows:

- Again, Isomap performs best for the training data, and also for testing data in low dimensions ($d < 8$).
- Isomap, PCA and MFCC performance reach a flat performance level from approximately 10 dimensions.
- A classification accuracy of 48.2% is achieved on the testing data with 5-dimensional Isomap features. This performance is only exceeded by much higher dimensional, ($d > 12$), MFCC and PCA features.

The mean classification accuracy results for each feature type in the ten vowel classification task are presented in Table 1. The mean accuracy scores were computed for the testing data evaluation. Averages are computed for three dimensionality ranges. It can be seen that Isomap gives the highest average accuracy in all ranges. LLE is shown to perform better than PCA and MFCC in low dimensions.

Dimensions	MFCC	PCA	Isomap	LLE
1–5	32.2000	38.3200	43.5600	38.6000
6–19	47.0000	47.0143	47.5286	44.6286
1–19	43.1053	44.7263	46.4842	43.0421

Table 1: Mean classification accuracy in the ten vowel classification task for MFCC, PCA, Isomap and LLE features.

Phone class classification results are presented in Fig. 3. The following is evident:

- Best accuracy is achieved in all dimensions with Isomap features.
- PCA and MFCC features yield similar performance, with PCA features offering improved accuracy for low dimensional features.

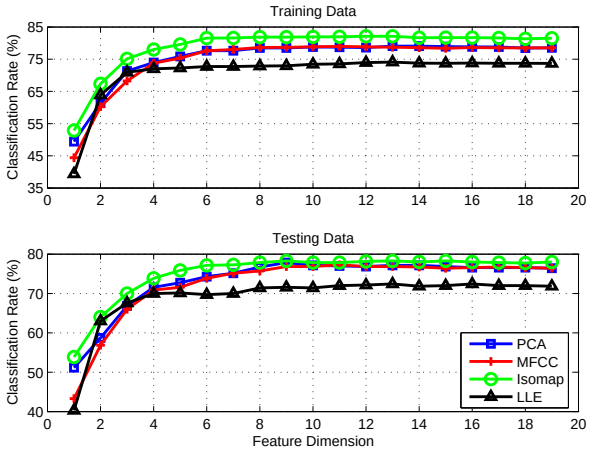

Figure 3: *Phone class classification results for baseline MFCC, PCA, Isomap and LLE features on the TIMIT database.*

- LLE features give the lowest classification rates, except for 2 and 3 dimensional features where they are second only to Isomap.

5. Conclusions

In this paper a phone classification system based on nonlinear manifold learning was proposed and evaluated against a baseline linear dimensionality reduction method, PCA, and conventional MFCC features. All of the dimensionality reduction methods presented outperform the baseline MFCC features for low dimensions. This illustrates the capability of these methods to extract discriminating information from the original 19-dimensional MFCC features.

Higher classification accuracy is shown for manifold learning derived features compared to baseline MFCC and PCA features for low dimensions. Also, in general Isomap yields superior performance to both MFCC and PCA features. This indicates that nonlinear manifold learning algorithms are more capable of retaining information required for discriminating between phones, especially in low dimensional space.

Comparing the manifold learning methods, Isomap demonstrates better classification accuracy than LLE. This indicates that preserving global structure rather than local relationships may be more important for speech feature transformation.

6. Future Work

Possible future work includes the application of the manifold learning feature transformation procedure presented here to continuous ASR. The manifold learning methods described above are batch processing algorithms. A number of out-of-sample extensions have been proposed to overcome this limitation. In the future these out-of-sample approaches could be developed for use with speech data.

7. Acknowledgments

Andrew Errity would like to acknowledge the support of the Irish Research Council for Science, Engineering and Technology; grant number RS/2003/114.

8. References

[1] W. Klein, R. Plomp, and L. C. W. Pols, "Vowel spectra, vowel spaces, and vowel identification," *J. Acoust. Soc. Amer.*, vol. 48, no. 4, pp. 999–1009, 1970.

[2] R. Togneri, M. Alder, and J. Attikiouzel, "Dimension and structure of the speech space," *IEE Proceedings-I*, vol. 139, no. 2, pp. 123–127, 1992.

[3] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 1–17, January 1999.

[4] A. Jansen and P. Niyogi, "Intrinsic Fourier analysis on the manifold of speech sounds," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[5] I. Jolliffe, *Principal Component Analysis*, ser. Springer Series in Statistics. New York: Springer-Verlag, 1986.

[6] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[7] X. Wang and K. Paliwal, "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition," *Pattern Recognition*, vol. 36, no. 10, pp. 2429–2439, October 2003.

[8] P. Somervuo, "Experiments with linear and nonlinear feature transformations in HMM based phone recognition based phone recognition," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, April 2003, pp. 52–55.

[9] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, December 2000.

[10] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[11] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in Neural Information Processing Systems*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14. Cambridge, MA: MIT Press, 2002, pp. 585–591.

[12] R. M. Hegde and H. A. Murthy, "Cluster and intrinsic dimensionality analysis of the modified group delay feature for speaker classification," *Lecture Notes in Computer Science*, vol. 3316, pp. 1172–1178, January 2004.

[13] V. Jain and L. Saul, "Exploratory analysis and visualization of speech and music by locally linear embedding," in *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2004, pp. 984–987.

[14] A. Errity and J. McKenna, "An investigation of manifold learning for speech analysis," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, Pittsburgh PA, USA, September 2006, pp. 2506–2509.

[15] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Machine Learning*, vol. 56, no. 1 - 3, pp. 209–239, July 2004.

[16] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, N.Y., 1995.

[17] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*, NIST, 1990.