A Comparison of Linear and Nonlinear Dimensionality Reduction Methods
Applied to Synthetic Speech

Andrew Errity and John McKenna

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

Abstract
In this study a number of linear and nonlinear dimensionality reduction methods are applied to high dimensional representations of synthetic speech to produce corresponding low dimensional embeddings. Several important characteristics of the synthetic speech, such as formant frequencies and f0, are known and controllable prior to dimensionality reduction. The degree to which these characteristics are retained after dimensionality reduction is examined in visualisation and classification experiments. Results of these experiments indicate that each method is capable of discovering meaningful low dimensional representations of synthetic speech and that the nonlinear methods may outperform linear methods in some cases.

Index Terms: dimensionality reduction, manifold learning, speech recognition, visualisation, synthetic speech.

1. Introduction
In the past, a variety of dimensionality reduction algorithms have been applied in a multitude of speech applications, including: feature transformation for improved speech recognition performance, speaker adaptation, speech compression, and speech analysis. The dimensionality reduction algorithms used have predominantly been linear and as such are limited to discovering the structure of data lying on or near a linear subspace of the high dimensional input space.

However, it has been proposed [1, 2] that speech sounds lie on a low dimensional manifold nonlinearly embedded in the high dimensional acoustic space. A low dimensional submanifold such as this may have a highly nonlinear structure that linear methods would fail to discover. A number of manifold learning (also referred to as nonlinear dimensionality reduction) algorithms have been proposed [3–5] which overcome the limitations of linear methods. These methods have been successfully applied to a number of benchmark manifold problems, image processing applications, and exploratory studies of speech, as summarised in previous studies [2, 6, 7].

In this study, we apply several manifold learning methods—locally linear embedding (LLE) [3], isometric feature mapping (Isomap) [4], and Laplacian eigenmaps (LEM) [5]—to synthetic speech data. The performance of these methods is also contrasted with that of the classical, linear, principal component analysis (PCA) method. In our previous work [6, 7] we have applied manifold learning methods to natural speech data. The motivation for using synthetic speech data in this study is that it facilitates the analysis of signals with known and controllable characteristics. Thus, after applying dimensionality reduction methods to these synthetic sounds we can examine, using both visualisation and classification experiments, the resulting lower dimensional data set to determine the degree to which these, known, characteristics have been retained.

The remainder of this paper is structured as follows. In Section 2, the speech synthesis technique used is outlined. Sections 3 and 4 describe the visualisation and classification experiments performed. Conclusions are presented in Section 5.

2. Synthetic speech generation
The speech synthesis technique used in this paper is similar to that described by Yegnanarayana et al. [8]. Synthetic speech was generated by exciting an LP-modelled filter with an artificially generated excitation signal. The excitation signal was generated using a classical LF-modelled glottal pulse train. This excitation signal was then applied to a set of 10 linear prediction coefficients, representing five formants, to produce a synthetic speech signal. The formant values used varied depending on the required experiment, as detailed in the following sections.

3. Visualisation

3.1. Introduction
Our initial experiment aimed to visually examine the ability of each dimensionality reduction method to discover low dimensional variation known to be present in a speech signal. A number of synthetic speech signals were generated in which important components of the speech signal were varied from the signal start to end. The purpose of introducing this variation is to determine if the dimensionality reduction methods can retain these important sources of variation in a lower dimensional embedding of the signal while discarding less relevant information. The three components varied included: the first and second formants (F1 and F2) and fundamental frequency (f0). These are three of the most important sources of information within a speech signal.

3.2. Data
Four types of synthetic speech signals were generated, they are described as follows:

- **Varying F1:** Initial F1 frequency of 300 Hz, increasing in equal sized increments reaching 700 Hz at the signal end. This resulted in a synthetic speech signal moving, approximately, from an /u/ to an /i/ sound in vowel space.

- **Varying F2:** Initial F2 frequency of 1000 Hz, increasing in equal sized increments reaching 2200 Hz at the signal end. This resulted in a synthetic speech signal moving, approximately, from an /u/ to an /i/ in vowel space.
3.3. Experiments

Each of the four artificially generated speech signals described above were analysed as follows. The speech signals were first pre-emphasised, with the filter $H(z) = 1 - 0.98z^{-1}$, and following this 13-dimensional MFCC feature vectors were computed from Hamming windowed 20 ms frames extracted with an overlap of 10 ms. This resulted in a data set of $N = 199$ MFCC feature vectors for each speech signal. Each of these data sets were then separately provided as input to the dimensionality reduction algorithms: Isomap, LLE, Laplacian eigenmaps, and PCA. The number of nearest neighbours, $k$, used in the manifold learning methods was chosen empirically. The value of $k$ was varied in the range $k = 2, \ldots, 30$ and the value which produced the best embeddings, as determined by visual inspection, was chosen for each of the manifold learning methods.

3.4. Results

The visualisation results are shown in Figures 1–3. The two dimensional visualisation of the signal with varying $F_1 & F_2$ is very similar to that shown in Figure 2; thus this visualisation is omitted. One can observe from Figures 1 and 2 that all four dimensionality reduction techniques discover the formant variation in the original speech signals in the two dimensional embedding spaces they produce. Notably, there is no clear difference in the clarity or extent to which the different techniques preserve the formant variation in the low dimensional space. It is also apparent, given the low dimensional embeddings of the speech signal with a varying $f_0$ trajectory shown in Figure 3, that all four dimensionality reduction methods are capable of retaining information relating to $f_0$ variation in a low dimensional space. However to provide clear visualisations of the variation in $f_0$, three dimensional embeddings—rather than two dimensional as in the case of formant variation—are required. It can also be seen that the variation in pitch is less well defined and separated than the formant variation shown in the Figures 1 and 2. This may indicate that more dimensions are required to accurately retain information relating to pitch variation than formant variation.

4. Classification of synthetic vowels

4.1. Introduction

In this second experiment we examined the ability of the dimensionality reduction methods to produce low dimensional repre-
This experiment involved performing two phone classification tasks.

4.2. Data

This experiment involved performing two phone classification tasks:

**Task 1** Five vowels: /a/, /i/, /u/, /e/, and /æ/.

**Task 2** Ten vowels: /a/, /i/, /u/, /e/, /æ/, /æ/, /o/, /o/, and /æ/.

Each of the ten vowel sounds listed above were synthesised using the technique described in Section 2. The frequencies of the first three formants used in the synthesis of each vowel, were based on findings presented in the classic study by Peterson and Barney [9]. The frequencies used are listed in Table 2. The fourth and fifth formants were kept fixed, as in Table 1. The formant bandwidths used are given in Section 3.2.

The classification of such well defined, noise free, spectrally consistent synthetic vowels would be a relatively simple task and would not reflect the difficulties associated with the classification of natural vowel sounds. In natural speech the formant values associated with vowel sounds, as listed in Table 2, are simply targets which the speaker attempts to reach but may in reality undershoot or overshoot due to factors such as coarticulation. As a result, we generated a set of synthetic vowel sounds in which the formant values where not fixed but varied slightly from one utterance to the next. This was accomplished by sampling the formant values from Gaussian distributions with means as indicated in Table 2 and with a standard deviation of 50 Hz for F1, 100 Hz for F2, and 250 Hz for F3. These deviations were chosen to give a reasonable range of variation for each formant.

Furthermore, in addition to the incorporation of formant variation we introduced a degree of noise corruption to the synthetic vowels in order to better approximate natural speech conditions. This involved adding a Gaussian white noise component, centred on the instant of glottal closure of each pitch period, as performed in a previous study [8].

Three levels of noise component duration were used: 0% (noise free), 60%, and 100% of the pitch period. The intensity of the noise used was varied in four levels, with signal-to-noise ratios (SNR) of: ∞ dB (no added noise), 20 dB, 10 db and 5 dB. Given the various combinations of noise duration and intensity, two separate sets of noise combinations were used: a low noise set and a high noise set, detailed in Table 3. For each synthetic vowel generated the type of noise corruption applied was selected at random from the possible combinations of either the low or high noise set. Separate experiments were performed on data from the low and high noise data sets.

4.3. Experiments

For each task 250 utterances of each vowel were synthesised, with formant and noise variation, as described above. Each synthesised utterance was 40 ms in length. Each utterance was preemphasised, as detailed in Section 3.3, and Hamming windowed. Following this, 13-dimensional MFCC feature vectors were computed for each windowed 40 ms frame.

The resulting feature vectors were concatenated to form an \( N \times D \) matrix, where \( N = 1250 \) in the five vowel task and \( N = 2500 \) in the ten vowel task. In both tasks \( D = 13 \). PCA, Isomap, LLE, and Laplacian eigenmaps were individually applied to reduce the dimensionality of this MFCC data matrix. The target dimensionality of these algorithms was varied, producing transformed feature vectors of dimensionality \( d = 1, \ldots, D \). The number of nearest neighbours, \( k \), used in the manifold learning methods was chosen empirically. The value of \( k \) was varied in the range \( k = 2, \ldots, 30 \) and the value which produced the highest classification accuracy was chosen for each of the manifold learning methods.

Phone classification experiments were performed using five different feature types: the original MFCC vectors and the features produced by applying PCA, Isomap, LLE, and Laplacian eigenmaps to the baseline MFCC vectors. For all feature types, a separate classification experiment, that is training and

---

**Table 2:** Formant frequencies (Hz) used for vowel synthesis.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>730</td>
<td>1090</td>
<td>2440</td>
</tr>
<tr>
<td>/i/</td>
<td>270</td>
<td>2290</td>
<td>3010</td>
</tr>
<tr>
<td>/u/</td>
<td>300</td>
<td>870</td>
<td>2240</td>
</tr>
<tr>
<td>/e/</td>
<td>530</td>
<td>1840</td>
<td>2480</td>
</tr>
<tr>
<td>/æ/</td>
<td>640</td>
<td>1190</td>
<td>2390</td>
</tr>
<tr>
<td>/æ/</td>
<td>390</td>
<td>1990</td>
<td>2550</td>
</tr>
<tr>
<td>/o/</td>
<td>500</td>
<td>1000</td>
<td>2100</td>
</tr>
<tr>
<td>/o/</td>
<td>450</td>
<td>1090</td>
<td>2300</td>
</tr>
<tr>
<td>/æ/</td>
<td>570</td>
<td>840</td>
<td>2410</td>
</tr>
<tr>
<td>/æ/</td>
<td>660</td>
<td>1720</td>
<td>2410</td>
</tr>
</tbody>
</table>

---

**Table 3:** Possible combinations of duration (%) and SNR (dB) of the noise components added to each pitch period of the synthetic vowel sounds.

<table>
<thead>
<tr>
<th>Duration</th>
<th>Possible SNR levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Low Noise Set</td>
</tr>
<tr>
<td>60</td>
<td>20,10</td>
</tr>
<tr>
<td>100</td>
<td>20,10</td>
</tr>
</tbody>
</table>
testing, was performed using feature vectors of dimensionality $d = 1, \ldots, D$. 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. The original MFCC vectors served as a baseline, also varying in dimensionality as detailed above.

Support vector machine (SVM) classifiers with radial basis function (RBF) kernels were used in these experiments. A more detailed discussion of the SVM classification procedure can be found in our previous work [7].

4.4. Results

The results of the low and high noise synthetic vowel classification experiments are exhibited in Figure 4. Comparing the performance of the various feature types, the following points can be observed from the results shown in Figure 4:

- In the majority of cases the dimensionality reduction methods outperform the baseline MFCCs in low dimensions, $d \leq 3$.
- While the locality preserving manifold learning methods, Laplacian eigenmaps and LLE, do not consistently outperform PCA or MFCC, the globally motivated Isomap does. Isomap embeddings were found to offer the best classification rate in the majority, 51.92%, of tests.
- The performance of the Laplacian eigenmaps algorithm is the most inconsistent of all the dimensionality reduction methods. The features produced by this algorithm frequently perform worse than the other features, including the baseline MFCCs; however Laplacian eigenmaps clearly yields the best overall performance in the high noise five vowel classification task.
- Interestingly LLE also performs well in the high noise five vowel classification task with an average classification rate second only to Laplacian eigenmaps. In fact the two local methods, LLE and Laplacian eigenmaps, offer similar performance in all four tests.
- The manifold learning methods perform particularly well in the high noise classification experiments, offering the best performance overall in this task.

5. Conclusions

As demonstrated in the visualisation experiments, each of the dimensionality reduction methods are clearly capable of discovering meaningful low dimensional representations of synthetic speech data. Results of synthetic phone classification show that the dimensionality reduction methods offer improved performance over baseline MFCCs, in very low dimensions ($d \leq 3$). For higher dimensions the dimensionality reduction methods were in general not found to offer a great improvement over the baseline MFCCs in the case of low noise; however in the case of high noise manifold learning methods were found to yield higher classification rates than MFCCs and PCA transformed features.

6. References