Dimensionality Analysis of Singing Speech Based on Locality Preserving Projections

Mahnoosh Mehrabani, John H. L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science, University of Texas at Dallas, USA
{mahmehrabani,john.hansen}@utdallas.edu

Abstract

In this study, we expand the question of "what is the intrinsic dimensionality of speech?" to "how does the intrinsic dimensionality of speech change from speaking to singing?". Our focus is on dimensionality of the vowel space regarding spectral features, which is important in acoustic modeling applications. Locality Preserving Projection (LPP) is applied for dimensionality reduction of the spectral feature vectors, and vowel classification performance is studied in low-dimensional subspaces. Performance analysis of singing and speaking vowel classification based on reducing the dimension shows that compared to speaking, a higher number of dimensions is required for effective representation of singing vowels. The results are also explained in terms of differences in the formant spaces of singing and speaking, and vowel classification performance is analyzed based on feature vectors consisting of formant frequencies. The formant analysis results are shown to be consistent with LPP dimensionality analysis, which verifies the inherent dimensionality increase of the vowel space from speaking to singing.
Index Terms: singing, dimensionality analysis, locality preserving projections

1. Introduction

Most speech applications are based on extracting feature vectors from short-time speech frames. The dimensionality of feature vectors varies for different applications. However, based on physiological constraints of speech production, the inherent dimensionality of speech is much lower than most feature vector dimensions. This study analyzes the dimensionality of singing speech and compares it to neutral speaking. The Locality Preserving Projections (LPP) subspace learning is used to study the underlying low-dimensional structure of singing and speaking.

Previous studies have shown that speech can efficiently be presented by a small number of parameters [1, 2, 3]. Dimensionality of speech has been analyzed, applying linear and nonlinear dimensionality reduction methods to spectral acoustic features, such as Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). Early studies [4, 5, 6] used Principal Component Analysis (PCA) to study linearly embedded low-dimensional manifold in the spectral feature space, and showed that the first two PCA dimensions accounted for most of the total variance in vowel space. They also showed that vowel representation in the two-dimensional PCA subspace of spectral features has a similar configuration to F2/F1 vowel plane, and the two spaces produce comparable vowel classification performance. Nonlinearly embedded low-dimensional subspaces have also been studied based on applying nonlinear manifold learning techniques [7, 8, 9].

Acoustic analysis of singing speech [10, 11, 12] indicates the deviation of vowel space from speaking to singing. However, little if any studies have explored the difference between intrinsic dimensionality of singing and speaking vowel spaces. We consider LPP dimensionality reduction [13] to study dimensionality of the vowel space for singing and compare it to speaking. LPP is a linear approximation of nonlinear manifold learning technique: Laplacian Eigenmap [14]. While PCA preserves the global structure of data, LPP preserves the neighborhood structure. The advantage of LPP over traditional PCA is that it can model nonlinear embedded manifold of data by preserving local relations among high-dimensional data points. In addition, though LPP shares many properties of nonlinear manifold learning techniques, unlike nonlinear methods that are defined just on the training data points, LPP is defined everywhere in ambient space and can be applied to unseen data.

In this study, the dimensionality of vowel spectra is analyzed and compared for singing and speaking based on vowel classification results in LPP subspaces of PLP feature vectors. It is shown that singing vowels require a higher number of dimensions than speaking to be efficiently represented. Furthermore, the formant space of singing and speaking vowels are compared, and vowel classification performances are presented based on formant frequency feature vectors. LPP dimensionality analysis of singing vowel space is shown to correlate with the formant space dimensionality, which verifies an increase in dimensionality of singing compared to speaking.

This study is based on a singing database, which includes singing and reading of lyrics for each speaker. Therefore, the phonetic contents of singing and speaking are the same. With the same speakers and the same text, the only changing factor is speaking style, and singing dimensionality can reliably be compared to speaking. The database is explained in more detail in the next section. LPP dimensionality reduction is described in Sec. 3. Sec. 4 represents vowel classification results in LPP subspaces of spectral feature vectors and dimensionality analysis. Formant analysis and vowel classification based on formant frequencies is presented in Sec. 5. Finally, conclusions are drawn in Sec. 6.

2. Database

Our experiments are based on a multilingual singing database (UT-Sing) [15], which includes singing and reading speech samples for each speaker with the same text. We collected UT-Sing for the purpose of singing speech analysis, comparing singing to speaking, and studying the effects of singing on various speech systems. Each speaker selected 5 popular songs in their native language. Each song was approximately 3-5 min-

utes in duration. The speaker's voice was recorded in a soundbooth with a close-talk microphone while singing as well as reading the lyrics of the same songs. Karaoke system prompts were used for singing. While subjects were listening to the music through headphones, the lyrics were displayed, and only the subjects singing voice was recorded (i.e., no music was captured within the audio stream).

For this study, four Hindi speakers, including two males and two females were selected based on their higher singing quality. For the remainder of this paper, we refer to the reading component of UT-Sing corpus as (neutral) speaking. Singing and speaking phonemes for all utterances from these four speakers were manually annotated by a trained transcriber fluent in Hindi. Table 1 shows total vowel counts for these speakers for the three most frequently used Hindi vowels in our database. The first row shows the International Phonetics Alphabet (IPA), and Devanagari symbols. The slight differences in the number of speaking and singing vowels are due to vowel insertions in singing. Our analysis is based on these three vowels since they had more than 60% of the total number of vowels in the phonetically transcribed utterances.

	/aː/ (आ)	/eː/ (ए)	/ə/ (अ)
Speaking	1441	1355	1343
Singing	1559	1469	1414

Table 1: *Hindi vowel counts.*

3. Locality Preserving Projections

Locality Preserving Projections (LPP) [13] is a linear unsupervised dimensionality reduction technique that optimally preserves the local neighborhood structure of the data. LPP is an alternative to Principal Component Analysis (PCA), a classical linear unsupervised dimensionality reduction process that projects the data along the directions with maximal variances. The observations in a high dimensional space, usually lie on a low dimensional manifold, and LPP and PCA seek the linearly embedded manifold in the data set. While PCA aims to preserve the global structure of the data set, LPP preserves the local structure. LPP has proven to outperform PCA in various applications [16, 17], including speaker clustering for singing speech [18]. LPP has also proven to be an effective feature transformation for speech recognition [19].

Given a set of n-dimensional data points: $x_1, \ldots, x_m$, a linear dimensionality reduction algorithm finds a transformation matrix A which maps these m data points to a set of vectors in an l-dimensional subspace: $y_1, \ldots, y_m$ such that $l << n$ and $y_i = A^T x_i, i = 1, \ldots, m$. LPP is in fact a linear approximation of the nonlinear Laplacian Eigenmap [14]. The LPP subspace learning algorithm first constructs an adjacency graph G with m nodes, where each node represents a data point. Two nodes i and j are connected if the corresponding data points x_i and x_j are "close". The concept of "closeness" of two data points is defined either in the sense of k nearest neighbor (i.e., i and j are connected if x_i is among k nearest neighbors of x_j and vice versa), or in the sense of ε-neighborhood (i.e., i and j are connected if $\|x_i - x_j\|^2 < \varepsilon$). Next, a weight is associated with each edge or each two connected nodes. The common weight function is the Heat Kernel:

$$W_{ij} = e^{-\|x_i - x_j\|^2 / t}, \quad (1)$$

where W is the weight matrix. Finally, the following objective function is minimized:

$$\sum_{ij} (y_i - y_j)^2 W_{ij} \quad (2)$$

Simple algebraic formulation [13] reduces the objective function to:

$$XLX^T a = \lambda XDX^T a \quad (3)$$

where $X = [x_1 \ldots x_m]$ is an $n \times m$ matrix of data vectors, D is a diagonal matrix such that: $D_{ii} = \sum_j W_{ij}$, and $L = D - W$ is the Laplacian matrix. Eq. (3) is a generalized eigenvalue problem, and the solutions $a_1, \ldots, a_l$ which are the eigenvectors ordered based on their corresponding eigenvalues are columns of an $n \times l$ matrix A such that:

$$y_i = A^T x_i, A = [a_1, \ldots, a_l]. \quad (4)$$

We calculated the LPP transformation matrix A for feature vectors extracted from singing and speaking vowel train sets, and applied this to reduce the dimension for vowel classification of test sets. More details are presented in the next section.

4. Dimensionality analysis

Our analysis of singing and speaking dimensions is based on vowel classification results in subspaces of the spectral feature space. Vowel classification was performed for the three vowels from Table 1, which had the most number of occurrences. For each of the four speakers, four songs were used for training, and one song for test. There was no overlap between train and test songs. First, in order to focus on sustained vowels and reduce the effect of coarticulation, speech frames were selected from the 50% middle of each vowel. Next, 12-dimensional PLP features were extracted from each frame. PLP feature vectors were classified using a k nearest neighbor classifier. Our initial experiments with full-dimensional feature vectors showed that increasing parameter k generally increases vowel classification accuracy for both speaking and singing, but for $k > 12$ the performance improvement is not significant. Therefore, parameter k was set to 12.

LPP dimension reduction was applied to feature vectors, and vowel classification was performed at the frame level with a decreasing number of dimensions: 12, 11, ..., 2, 1. Fig. 1 shows vowel classification accuracy for each dimension. As shown, speaking and singing have approximately the same classification accuracies with full-dimension PLP features. This can be interpreted as similar vowel separability for these three vowels with full-dimensional PLPs for speaking and singing. However, singing vowels are expected to have more variability than speaking. Therefore, it is hypothesized that a higher number of dimensions is required to efficiently represent singing vowels.

Fig. 1 verifies this hypothesis. From dimension 11 to 4, both speaking and singing have similar vowel classification performance to the baseline (full dimension) with standard deviation of 0.5. However, from dimension 4 to 3, singing vowel classification accuracy decreases by 9.4%, while the relative accuracy decrease for speaking is 1.8%. With only two dimensions, speaking vowel classification performance is similar to that of the baseline, and there is a relative performance loss of

27.7% when decreasing the dimension from 2 to 1. This implies that the first two dimensions can efficiently represent these three vowels for speaking, yet for singing vowels at least four dimensions are required. In order to visualize this inherent dimensionality difference between speaking and singing, scatter plots of 3-dimensional feature vector projections for the most separable vowel pair in this vowel set are depicted in Fig. 2. As shown, speaking feature vectors are separable even if projected on a 2-dimensional plane, while for singing more than three dimensions is required to separate feature vectors for these vowels.

To compare LPP dimensionality reduction to traditional PCA for vowel classification, Fig. 3 illustrates the singing vowel classification performance when reducing the dimension from 12 to 1 for LPP and PCA subspaces. As shown, PCA has worse performance compared with LPP for almost all dimensions with an average performance loss of 5%. Unlike LPP, PCA does not have consistent performance for dimensions 12 to 4, and the classification accuracy fluctuates with a standard deviation of 2.1. In the next section, we will show that vowel classification results based on formant frequencies correlate with LPP results, which implies the nonlinearity of an embedded singing vowel space. As noted, in this study we use LPP as a linear approximation of nonlinear manifold learning to apply the transformation matrix trained with training vowels to the unseen vowel test set. The next section explains dimensionality analysis results in terms of formant space analysis.

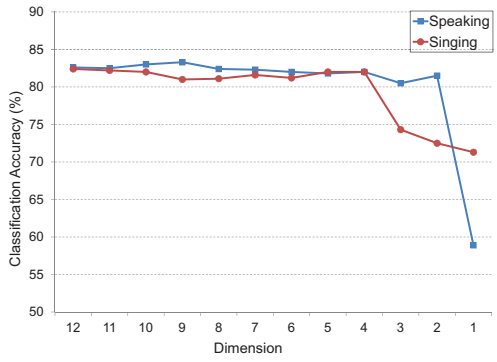


Figure 1: *Vowel classification results for speaking and singing when reducing the dimension from 12 to 1 in LPP subspace.*

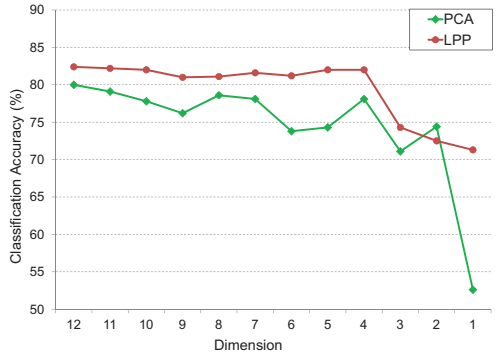


Figure 3: *Vowel classification results for singing when reducing the dimension from 12 to 1 in LPP and PCA subspaces.*

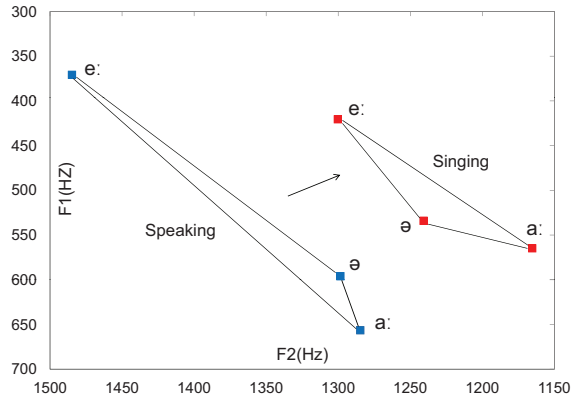


Figure 4: *Transformation of F2/F1 vowel configuration from speaking to singing.*

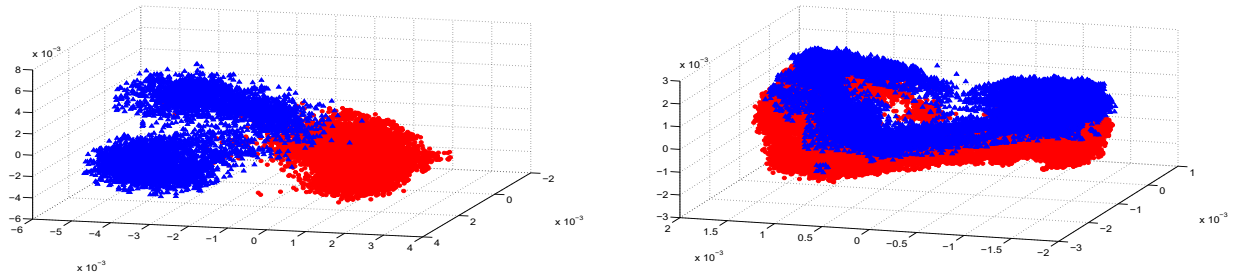
5. Formant analysis

Acoustic analysis of singing has shown the spectral deviation of singing vowels from speaking due to articulatory modification while singing [20, 21, 22]. Those studies verify the changes in formant frequencies of sung vowels compared to spoken vowels. While most of the previous studies analyzed isolated sung vowels, a recent study [12] showed the variations between singing and speaking vowel spaces in context.

We studied formant frequencies for the three Hindi vowels from Table 1 and related the formant space changes in the singing vowel space compared to speaking, to dimensionality differences between singing and speaking. For each vowel, four formant frequencies were extracted with an LPC order = 12, interval length = 0.01 sec. and analysis window length = 0.05 sec. Formant frequencies were estimated for the 50% middle of vowels that had duration more than 0.1 sec. Fig. 4 illustrates how the vowel configuration in F2/F1 plane changes from speaking to singing. The presented F2/F1 configuration is based on mean formant frequencies. Though the distance between the two most confusable vowels in this vowel set has increased, the average Euclidean distance in F2/F1 plane between vowels has been reduced by 36.9% from speaking to singing. This helps explain why with two dimensions, speaking vowel classification has much higher accuracy than for singing. Next, vowel classification was performed using formant frequencies as feature vectors with the same train and test sets applied for dimensionality analysis in Sec. 4. For formant based singing versus speaking dimensionality analysis, the dimension reduction was achieved by dropping higher order formants first, which is shown to produce similar results to LPP dimension reduction. Table 2 summarizes the results with formant vector dimensions: 4:$[F1, F2, F3, F4]$, 3:$[F1, F2, F3]$, 2:$[F1, F2]$, and 1:$[F1]$.

Dimension	*4*	*3*	*2*	*1*
Speaking	88.9%	88.4%	82.0%	67.6%
Singing	82.6%	80.2%	68.2%	61.2%

Table 2: *Vowel classification results for speaking and singing using formant frequency features when reducing the dimension from 4 to 1.*



(a) (b)

Figure 2: *3-dimensional feature vector scatter plot of two vowels (one in blue, one in red) for (a) speaking and (b) singing.*

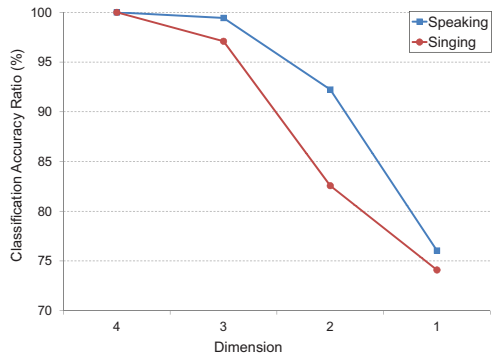


Figure 5: *Accuracy ratio (%) of vowel classification using 4, 3, 2, and 1 formants to vowel classification using 4 formants for speaking and singing.*

With formant frequencies as feature vectors, singing vowel classification has always lower performance than speaking. However, the maximum performance loss from speaking to singing occurs at dimension 2 (i.e., using the first two formant frequencies). This shows that speaking vowels are much more separable than singing vowels with the first two formants as the only two dimensions. Fig. 5 depicts relative classification accuracies (i.e., classification accuracy at each dimension $\times$ 100 divided by maximum classification accuracy with four formants). As shown, with the first two formants, for speaking 92.2% of performance with four formants is achieved, while for singing 82.6% of performance is obtained. This result is consistent with formant configuration analysis in Fig. 4. In addition, it correlates with LPP dimensionality analysis in Fig. 1, which confirms the intrinsic dimensionality increase of vowel space from speaking to singing.

6. Conclusions

Singing vowel space variation from speaking was studied in terms of dimensionality analysis. The hypothesis that singing vowels require more dimensions than neutral speaking for efficient representation was verified based on vowel separability analysis by reducing the dimension of spectral feature vectors. LPP subspace learning was applied to represent low-dimensional manifolds, while preserving neighborhood structure of the data. It was shown that for speaking with two LPP dimensions, approximately 99% of full-dimensional vowel classification accuracy was achieved. However, for singing 88% of full-dimensional classification performance was obtained using 2-dimensional LPP feature vectors. A similar result for vowel classification performance with the first two formant frequencies, confirmed the higher intrinsic dimensionality of singing vowel space compared to speaking. The results were also explained based on different configurations of speaking and singing vowels in the formant space.

This study was a first attempt to analyze dimensionality of singing speech. It was shown that for low-dimensional representation, singing requires more dimensions than speaking. This result can be applied to acoustic modeling of singing speech for various applications, such as speaker and language classification for singing, and singing speech recognition and phoneme alignment. Due to the lack of transcribed singing speech and acoustic models for singing, and for more reliable results phonemes were manually annotated. Therefore, the experiments were conducted for a limited number of speakers, and vowels with enough number of occurrences for statistical analysis. However, the reliability of results are based on the same phonetic context for singing and speaking, as well as having 5 songs per speaker for a song independent analysis. Future research includes analyzing the dimensionality of singing for more languages, and comparing the results, and applying dimensionality reduction to a larger set of vowels with a wider variety of feature vectors.

7. References

[1] M. Alder, R. Togneri, and Y. Attikiouzel, "Dimension of the speech space," in *Communications, Speech and Vision, IEE Proceedings I*, vol. 138, no. 3, 1991, pp. 207–214.

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

[3] A. Errity, "Exploring the dimensionality of speech using manifold learning and dimensionality reduction methods," Ph.D. dissertation, Dublin City University, 2010.

[4] R. Plomp, L. Pols, and J. Van der Geer, "Dimensional analysis of vowel spectra," *Journal of Acoustical Society of America*, vol. 41, pp. 707–712, 1967.

[5] W. Klein, R. Plomp, and L. C. Pols, "Vowel spectra, vowel spaces, and vowel identification," *Journal of the Acoustical Society of America*, vol. 48, no. 4B, pp. 999–1009, 1970.

[6] L. C. Pols, H. R. Tromp, and R. Plomp, "Frequency analysis of Dutch vowels from 50 male speakers," *Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1093–1101, 1973.

[7] A. Errity and J. McKenna, "An investigation of manifold learning for speech analysis," in *Interspeech*, 2006, pp. 2506–2509.

[8] A. Errity, J. McKenna, and B. Kirkpatrick, "Manifold learning-based feature transformation for phone classification," *Advances in Nonlinear Speech Processing*, pp. 132–141, 2007.

[9] R. Hegde and H. Murthy, "Cluster and intrinsic dimensionality analysis of the modified group delay feature for speaker classification," in *Neural Information Processing*, 2004, pp. 1172–1178.

[10] J. Sundberg, *The acoustics of the singing voice*. Scientific American, 1977.

[11] I. R. Titze, "Speaking vowels versus singing vowels," *Journal of Singing*, vol. 52, no. 1, pp. 41–42, 1995.

[12] E. D. Bradley, "An investigation of the acoustic vowel space of singing," in *Proceedings of 11th International Conference on Music Perception and Cognition*, 2010.

[13] X. He and P. Niyogi, "Locality preserving projections," in *Proceedings of Advances in Neural Information Processing Systems*, vol. 16, 2003, p. 153.

[14] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in neural information processing systems*, vol. 14, pp. 585–591, 2002.

[15] M. Mehrabani and J. H. Hansen, "Language identification for singing," in *IEEE ICASSP*, 2011, pp. 4408–4411.

[16] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacianfaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 328–340, 2005.

[17] S. Chu, H. Tang, and T. Huang, "Locality preserving speaker clustering," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 2009, pp. 494–497.

[18] M. Mehrabani and J. Hansen, "Singing speaker clustering based on subspace learning in the GMM mean supervector space," *Speech Communications*, vol. 55, no. 5, pp. 653–666, 2013.

[19] Y. Tang and R. Rose, "A study of using locality preserving projections for feature extraction in speech recognition," in *IEEE ICASSP*, 2008, pp. 1569–1572.

[20] G. Bloothooft and R. Plomp, "Spectral analysis of sung vowels. i. variation due to differences between vowels, singers, and modes of singing," *Journal of the Acoustical Society of America*, vol. 75, p. 1259, 1984.

[21] J. Sundberg, "The science of the singing voice," *Northern Illinois University Press*, 1987.

[22] N. Henrich, J. Smith, and J. Wolfe, "Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones," *Journal of the Acoustical Society of America*, vol. 129, p. 1024, 2011.