Learning One-to-Many Mapping With Locally Linear Maps Based on Manifold Structure

Do-Kwan Oh, Sang-Hoon Oh, and Soo-Young Lee

***Abstract*—This letter proposes a new method to realize a nonlinear mapping of one-to-many correspondences. Assuming that a small number of training pairs are given with their actual correspondences, each tangent space is locally constructed on a submanifold around each labeled sample. Moreover, the linear transformation between paired tangent spaces is derived by solving an optimization problem, which is designed to bring locally linear maps into closer proximity in each class. Finally, a global nonlinear mapping is realized by combining these locally linear maps. In simulations of an S-curve to Swiss-roll, a lip to speech, and room impulse response to position of microphone mappings, the proposed method shows the remarkable mapping ability.**

***Index Terms*—Lip reading, lip-to-speech mapping, manifold learning, monaural source localization, one-to-many correspondence.**

I. Introduction

In most cases, human sensory data resides in very high dimensional spaces; however, the underlying systems that generate this data may have relatively low degrees of freedom. From this hypothesis, it has been assumed that human sensory data is likely to lie on a manifold embedded in a high-dimensional space. Many nonlinear dimension reduction algorithms based on the common perspective about the structure of human sensory data, including LLE [1], ISOMAP [2], and LTSA [3], have been proposed. This manifold concept was applied to learn the locally linear map between two image datasets with one-to-one matched pose variations [4].

In real environments, multimodal datasets such as lip images and speech signals from a camcorder assume a one-to-one correspondence along time. However, human can make various sounds with a similar lip shape because the lip is only one part among lots of articulators in the speech production mechanism. Therefore the general mapping from the space of visemes to the space of phonemes has a one-to-many correspondence. Also, there are lots of possibilities of one-to-many correspondences in other datasets. Thus, a one-to-many mapping algorithm should be realized.

This letter proposes a new method of one-to-many mapping by combining a clustering scheme with a one-to-one mapping scheme [4]. The contributions of this letter are the realization of mapping between two datasets with a one-to-many correspondence and the demonstration of two real life application.

Section II introduces a *manifold-constrained map* [4] and the overall mapping structure proposed in this letter. Section III defines an optimization problem with a clustering scheme. Section IV presents three experiments and their results. Finally, Section V provides the conclusion.

II. Manifold-Constrained Map

Let $\boldsymbol{X} = \{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbb{R}^{d_x}; i = 1, 2, \cdots, n_x\}$ and $\boldsymbol{Y} = \{\boldsymbol{y}_j | \boldsymbol{y}_j \in \mathbb{R}^{d_y}; j = 1, 2, \cdots, n_y\}$ be the two different types of training datasets. Among the training datasets, n pairs of matched data $(\boldsymbol{u}_k, \boldsymbol{v}_k)$ are given with their actual correspondence: $\boldsymbol{U} = \{\boldsymbol{u}_k | \boldsymbol{u}_k \in \mathbb{R}^{d_x}; k = 1, 2, \cdots, n\}$ and $\boldsymbol{V} = \{\boldsymbol{v}_k | \boldsymbol{v}_k \in \mathbb{R}^{d_y}; k = 1, 2, \cdots, n\}$, where $n \ll n_x$ and $n \ll n_y$. At each labeled sample $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$, their corresponding local tangent spaces with basis matrices $\boldsymbol{S}_k$ of rank $d_S (\ll d_x)$ and $\boldsymbol{T}_k$ of rank $d_T (\ll d_y)$ are constructed on the lower-dimensional submanifolds, which include the unlabeled training samples surrounding $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$, respectively. Our ultimate objective is to determine linear transformations $\boldsymbol{L}_k$ between the paired kth local tangent spaces. Therefore, a *locally linear map* [4] is defined as

$$\boldsymbol{f}_k(\boldsymbol{x}) = \boldsymbol{T}_k{}^T \boldsymbol{L}_k \boldsymbol{S}_k (\boldsymbol{x} - \boldsymbol{u}_k) + \boldsymbol{v}_k \tag{1}$$

which is linearly transformed along the pair of kth local tangent spaces. The inverse map, $\boldsymbol{T}_k^T$, is necessary for recovering the local coordinates to the global coordinate. Most manifold learning techniques have a common drawback, which is hard to find an inverse map. This is the reason why we substitute the simple linear tangent space for nonlinear dimension reduction technique to represent each local geometry.

A global nonlinear mapping function [4] is approximated by combining the locally linear maps of (1) as

$$\boldsymbol{f}(\boldsymbol{x}_i) \approx \sum_k \alpha_{ik} \boldsymbol{f}_k(\boldsymbol{x}_i). \tag{2}$$

Weight α_{ik} is responsible for the contribution of the kth linear map $\boldsymbol{f}_k(\boldsymbol{x}_i)$ to $\boldsymbol{f}(\boldsymbol{x}_i)$. Because it is difficult to evaluate the exact value of α_{ik}, we use a value that is inversely proportional to the distance between $\boldsymbol{x}_i$ and $\boldsymbol{u}_k$.

Fig. 1 shows the proposed mapping structure and notations. Five linear maps are exemplified and two candidates for the final mapping output are illustrated with two colored classes which

Manuscript received May 07, 2011; revised June 14, 2011; accepted June 17, 2011. Date of current version July 19, 2011. This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) through the Ministry of Education, Science and Technology (2009-0092812). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. H. Vicky Zhao.

D.-K. Oh is with the Department of Bio and Brain Engineering, KAIST, Daejeon 305-701, Korea (e-mail: origell@neuron.kaist.ac.kr).

S.-H. Oh is with the Department of Information Communication Engineering, Mokwon University, Daejeon, Korea (e-mail: shoh@mokwon.ac.kr).

S.-Y. Lee is with the Department of Electrical Engineering, KAIST, Daejeon 305-701, Republic of Korea (e-mail: sylee@kaist.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2011.2161578

1070-9908/$26.00 © 2011 IEEE

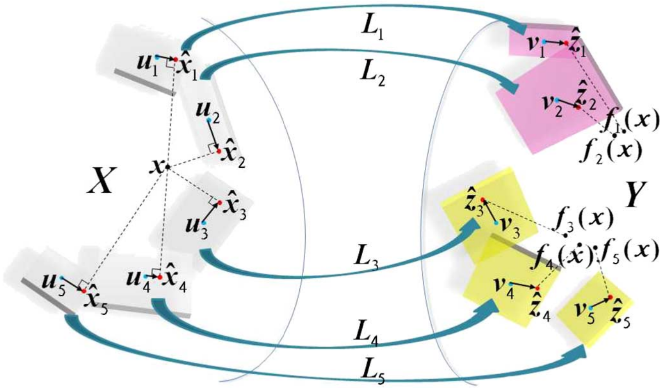

Fig. 1. Overall mapping structure.

are assigned from a cluster scheme. A detailed discussion of the clustering scheme is provided in Section III.

III. Optimization Problem

The most natural design for learning $\boldsymbol{L}_k$ is to minimize the discrepancy of each linear map that belongs to the same class. In this sense, the cost function is proposed as

$$E(\boldsymbol{L}) = \sum_{i=1}^{n_x} \sum_{m \in \boldsymbol{C}(\boldsymbol{x}_i)} \left[tr \left\{ \sum_{k \in \boldsymbol{G}_i(m)} \alpha_{ik} \boldsymbol{f}_k(\boldsymbol{x}_i) \boldsymbol{f}_k(\boldsymbol{x}_i)^T \right. \right.$$
$$\left. \left. - \sum_{k \in \boldsymbol{G}_i(m)} \alpha_{ik} \boldsymbol{f}_k(\boldsymbol{x}_i) \sum_{j \in \boldsymbol{G}_i(m)} \alpha_{ij} \boldsymbol{f}_j(\boldsymbol{x}_i)^T \right\} \right] \quad (3)$$

which is a trace norm of the weighted covariance of different linear maps in the same class. Here, $\boldsymbol{L}$ is a set of $\boldsymbol{L}_k$ for $k = 1, 2, \ldots, n$. In addition, $\boldsymbol{C}(\boldsymbol{x}_i)$ denotes a set of class indices of tangent spaces that contain $\boldsymbol{x}_i$, and $\boldsymbol{G}_i(m)$ denotes a set of tangent space indices involved with the mth class. Thus, $\sum_{k \in \boldsymbol{G}_i(m)} \alpha_{ik}$ is normalized to one for each $m \in \boldsymbol{C}(\boldsymbol{x}_i)$.

Contrary to the cost function of [4], (3) adopts a clustering scheme for one-to-many mapping. This is the main difference between [4] and the proposed method. The class information $\boldsymbol{C}(\boldsymbol{x}_i)$ in (3) is given by the k-*means* algorithm on a *geodesic* metric driven from the entire training data $\boldsymbol{Y}$. In the presence of a metric, *geodesics* are defined to be (locally) the shortest path between points in the space.

For modeling the distribution and the size of each tangent space, two initial metric matrices are designed using parameters K and P respectively. K gives the number of training samples that belong to one tangent space and P describes the number of tangent spaces that are involved with an input sample. The final metric matrix is computed by superimposing the previous two metric matrices.

We can find the minimum of (3) with the condition

$$\frac{\partial E(\boldsymbol{L})}{\partial \boldsymbol{L}_k} = 0, \quad \text{for } k = 1, 2, \cdots, n. \quad (4)$$

From (3) and (4), n linear matrix equations are given by

$$\boldsymbol{T}_k \boldsymbol{T}_k^T \boldsymbol{L}_k \boldsymbol{S}_k \boldsymbol{A}_k \boldsymbol{S}_k^T + \boldsymbol{T}_k \boldsymbol{B}_k \boldsymbol{S}_k^T - \sum_{j \in \boldsymbol{I}_k} \boldsymbol{T}_k \boldsymbol{T}_j^T \boldsymbol{L}_j \boldsymbol{S}_j \boldsymbol{C}_{jk} \boldsymbol{S}_k^T -$$
$$\boldsymbol{T}_k \boldsymbol{D}_k \boldsymbol{S}_k^T = 0, \quad \text{for } k = 1, 2, \cdots, n \quad (5)$$

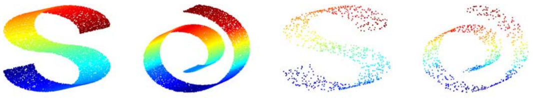

Fig. 2. Training data for the S-curve and Swiss-roll mapping. (a) X; (b) Y; (c) U; (d) V.

where $\boldsymbol{I}_k$ denotes the index of tangent spaces overlapped with the kth tangent space in the same class. Matrices $\boldsymbol{A}_k$, $\boldsymbol{B}_k$, $\boldsymbol{C}_{jk}$, and $\boldsymbol{D}_k$ are derived from the data itself as

$$\boldsymbol{A}_k = \sum_i \alpha_{ik} (\boldsymbol{x}_i - \boldsymbol{u}_k)(\boldsymbol{x}_i - \boldsymbol{u}_k)^T,$$
$$\boldsymbol{B}_k = \sum_i \alpha_{ik} \boldsymbol{v}_k (\boldsymbol{x}_i - \boldsymbol{u}_k)^T,$$
$$\boldsymbol{C}_{jk} = \sum_i \alpha_{ij} \alpha_{ik} (\boldsymbol{x}_i - \boldsymbol{u}_j)(\boldsymbol{x}_i - \boldsymbol{u}_k)^T,$$
$$\boldsymbol{D}_k = \sum_i \left(\sum_{j \in \boldsymbol{I}_k} \alpha_{ij} \boldsymbol{v}_j \right) \alpha_{ik} (\boldsymbol{x}_i - \boldsymbol{u}_k)^T. \quad (6)$$

Thus, (5) can be solved since it consists of n linear matrix equations with n unknown matrices $\boldsymbol{L}_k$, $k = 1, 2, \cdots, n$. In this letter, the left matrix division operator, *mldivide* \, is used as a linear equation solver in MATLAB simulations.

In the proposed method, each tangent space has a unique rank. Even if the initial rank is sufficiently large, the coefficients corresponding to superfluous intrinsic coordinates are trained as close to zero in a linear transformation matrix. Thus, the proposed algorithm also learns an effective relationship between paired tangent spaces.

IV. Experiments

A. S-Curve to Swiss-Roll Mapping

The first experiment is an *S-curve* to *Swiss-roll* mapping with artificial data. We consider two types of training datasets, namely, $\boldsymbol{X}$ [Fig. 2(a)] and $\boldsymbol{Y}$ [Fig. 2(b)]. The color coding information shows the correspondence between $\boldsymbol{X}$ and $\boldsymbol{Y}$. It is noteworthy that the *Swiss-roll* is color coded by height, not rotation. Thus, the correspondence between the *S-curve* and the *Swiss-roll* is maximally one-to-three and minimally one-to-one. The number of samples in $\boldsymbol{X}$ and $\boldsymbol{Y}$ is 16,000 $(= n_x, n_y)$, and their dimension is set to 3 $(= d_x, d_y)$. In this simulation, only 5% of $\boldsymbol{X}$ are randomly selected as $\boldsymbol{U}$ [Fig. 2(c)], and the corresponding $\boldsymbol{Y}$ are chosen as $\boldsymbol{V}$ [Fig. 2(d)].

Fig. 3(a) shows 4000 test samples of *Swiss-roll* treated as the target for the mapping output. Without applying any clustering scheme, as shown in Fig. 3(b), the boundary between each class will disappear and this blurring effect between classes will destroy the entire data structure. Fig. 3(b) also coincides with the result of [4]. This is the main problem that must be resolved for one-to-many mapping and the reason why a clustering technique must be inevitably applied.

The proposed method uses the locally linear map defined as (1). On the contrary, we adopt a baseline method which uses only the labeled data as $\boldsymbol{f}_k(\boldsymbol{x}) = \boldsymbol{v}_k$. All the other parameters are identical to those in the proposed algorithm. This baseline is the most appropriate for the verification of the mapping algorithm. Fig. 3(c) is the simulation result for the baseline.

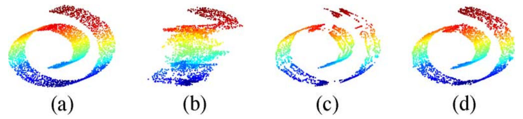

Fig. 3. Test results for the S-curve to Swiss-roll mapping. (a) Target; (b) without cluster; (c) baseline; (d) proposed.

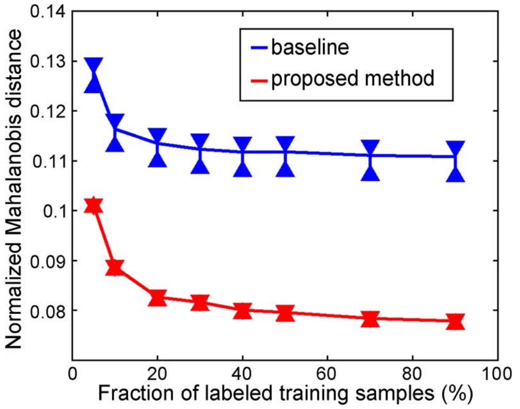

Fig. 4. Distance error for the S-curve to Swiss-roll mapping.

Fig. 3(d) shows the result of the proposed method, which appears denser than Fig. 3(c). A mapping output is defined as (2). Thus, the possible mapping position of the baseline, $\boldsymbol{f}(\boldsymbol{x}_i) \approx \sum_k \alpha_{ik}\boldsymbol{v}_k$, is limited within an area composed of the selected samples $\boldsymbol{v}_k$'s. As a result, the baseline method does not have mapping outputs in the inter-class area. However the mapping output of the proposed method can be more freely located on well-distributed tangent spaces by the $\boldsymbol{T}_k^T\boldsymbol{L}_k\boldsymbol{S}_k(\boldsymbol{x}-\boldsymbol{u}_k)$ term.

The number of clusters could be three due to the maximal one-to-three correspondence. However it is hard to get ideal cluster boundaries exactly by an automatic clustering technique. We ascertained that the ideal boundary information is successfully included when k of the k-*means* is 20. Also, the maximum rank on each tangent space is set to 3. In the test phase, we only need to reevaluate α_{ik} for each test sample. The closest candidate to the target becomes the final mapping output.

Under the assumption that elements of feature vectors are uncorrelated, we define the *Normalized Mahalanobis distance* by $d_{NM}(\boldsymbol{x},\boldsymbol{y}) = \sqrt{(1/N)(\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})}$, where $\boldsymbol{\Sigma}$ is a diagonal matrix whose main diagonal entries are the variances of each element over the whole training sample set and N is the feature dimension. Fig. 4 shows the average of the distances between the mapping outputs and targets for all test samples. The x-axis represents the fraction of the labeled training samples. The performance of the proposed method is better than that of the baseline. It is noteworthy that the final mapping output is closer to the target, although we do not use any target information during the training procedure.

In the simulation, α_{ik} is calculated by $e^{-\|\boldsymbol{x}_i-\boldsymbol{u}_k\|/\sigma}$. In Fig. 4, each solid curve is drawn from $\sigma = 1$, and the upper and lower triangular bars are drawn from $\sigma = 2$ and $\sigma = 0.5$, respectively. Because the proposed algorithm is able to minimize the discrepancy of each linear map $\boldsymbol{f}_k(\boldsymbol{x}_i)$, our method is more robust to α_{ik} than the baseline method.

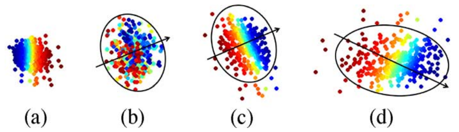

Fig. 5. Locally linear map from a lip tangent space to a speech tangent space. (a) $\boldsymbol{S}_{75}(\boldsymbol{x}_i^{(75)}-\boldsymbol{u}_{75})$. (b) $\boldsymbol{T}_{75}(\boldsymbol{y}_i^{(75)}-\boldsymbol{v}_{75})$. (c) $\boldsymbol{L}_{75}\boldsymbol{S}_{75}(\boldsymbol{x}_i^{(75)}-\boldsymbol{u}_{75})$ using $\boldsymbol{L}_{75}$ of the proposed method. (d) $\boldsymbol{L}_{75}\boldsymbol{S}_{75}(\boldsymbol{x}_i^{(75)}-\boldsymbol{u}_{75})$ using $\boldsymbol{L}_{75}$ of the without clustering method.

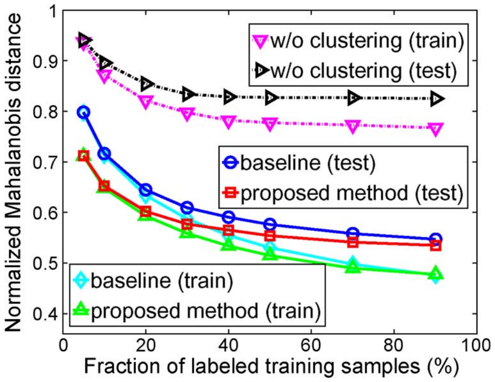

Fig. 6. Distance error for the lip-to-speech mapping.

The parameters K and P, which were introduced in the Section III, are related to local geometries of training samples. K is optimized from the reconstruction error defined as (see the bottom of the page), where $\boldsymbol{x}_i^{(k)}$ denote the ith samples on the kth tangent space and $n_x^{(k)}$ denote the number of $\boldsymbol{x}_i^{(k)}$'s. The averages of E_k^{RC} over all k are minimized at $K = 400$ in the two artificial data spaces and two real data spaces (lip and speech), respectively. To simplify the modeling of local geometry, P is used only for the compensation of outlier samples which do not have any neighbor labeled frame. Therefore P is set to 1.

B. Lip-to-Speech Mapping

For more verification in real applications, we simulated the lip-to-speech mapping using the *CUAVE* database [5]. The dataset used in this experiment consists of the ten English digits recorded five times by 36 speakers, and one record among the five repetitions is used as a test set. The number of training frames is 17 344 ($= n_x, n_y$), and the number of test frames is 4396. The lip feature is 20 delta features and 20 acceleration features from 20 static PCA coefficients extracted from raw pixel (30 × 70) vectors. The speech feature is the 36 Mel Frequency Cepstrum Coefficients. In this experiment, the maximum rank on each tangent space is set to 5 and the number of cluster is set to 60 based on the number of pronouncing English phonemes.

Fig. 5 shows an example of locally linear maps from a lip tangent space to a speech tangent space. Let $\boldsymbol{x}_i^{(75)}$ denote the feature samples which are involved on the 75th tangent space in the lip dataset and $\boldsymbol{y}_i^{(75)}$ denote the feature samples corresponding to $\boldsymbol{x}_i^{(75)}$ in the speech dataset. Fig. 5(a) shows the projection of $\boldsymbol{x}_i^{(75)}$ onto 2-D 75th lip tangent space. These samples are color coded by left to right. Fig. 5(b) shows the projection of $\boldsymbol{y}_i^{(75)}$

$$E_k^{RC} = (1/n_x^{(k)})\sum_{i=1}^{n_x^{(k)}}\sqrt{(1/N)\|\{(\boldsymbol{x}_i^{(k)}-\boldsymbol{u}_k)-\boldsymbol{S}_k^T\boldsymbol{S}_k(\boldsymbol{x}_i^{(k)}-\boldsymbol{u}_k)\}^T\boldsymbol{\Sigma}^{-1}\|}$$

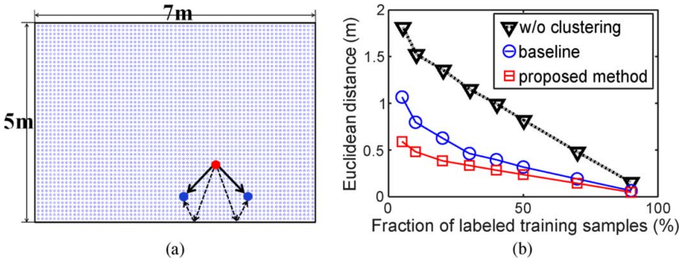

Fig. 7. Experimental environment and results for monaural source localization. (a) At two different positions of blue dots, RIRs may be similar, because a direct path and the first reflection path are same. (b) Distance error for the monaural source localization.

onto 2-D 75th speech tangent space, which are treated as target samples. Fig. 5(c) and 5(d) show the samples linearly transformed by $\boldsymbol{L}_k$ of the proposed and without clustering methods, respectively.

In Fig. 5(b)–(d), each ellipse shows the outline of tangent space and each arrow means the direction of color coding from red to blue, which show the correspondence to Fig. 5(a). From this experiment, we confirm that the proposed method can learn the linear transformation (rotation angle and scale) better than the without clustering method [4]. A correspondence error measure can be also defined as $E_k^{CP} = (1/n_x^{(k)}) \sum_{i=1}^{n_x^{(k)}} \sqrt{(1/N)\|\{\boldsymbol{T}_k(\boldsymbol{y}_i^{(k)} - \boldsymbol{v}_k) - \boldsymbol{L}_k \boldsymbol{S}_k(\boldsymbol{x}_i^{(k)} - \boldsymbol{u}_k)\}^T \boldsymbol{\Sigma}^{-1}\|}$, where $\boldsymbol{y}_i^{(k)}$ denote the samples corresponding to $\boldsymbol{x}_i^{(k)}$ in speech dataset. From the fact that the correspondence errors of Fig. 5(c) and 5(d) are 0.4003 and 0.7689, we can argue that the proposed method is still better in numerical comparison.

Fig. 6 shows the mean of the normalized Mahalanobis distance between the mapping outputs and the targets for all unlabeled training samples and all test samples, respectively. In this figure, the dotted curves show the performances of one-to-one mapping [4] without the clustering scheme. Even though this is a real problem, the proposed method still shows better mapping performances than the other methods.

If we assume a lattice with a spacing of $1/R$ between adjacent points, an equivalent sampling of a 36-D unit hypercube would require R^{36} sample points. Since we have 17 344 training samples, the interval $(1/R)$ of evenly spaced samples in the speech space is around 0.7625 by the calculation of $R^{36} = 17\,344$. Thus we can argue that a mapping output is located near the target within a sample interval radius.

C. Monaural Source Localization

To demonstrate the benefit of the proposed method, we simulated one more real application, a *Room Impulse Response* (RIR) to *Position of Microphone* mapping. The relationship between these two datasets is nonlinear, because sound reflection mechanism has a nonlinear operation. Moreover, symmetric rectangular parallelepiped room structure causes one-to-many correspondence between RIRs and positions.

The size of the recording room is $7 \times 5 \times 2.75$ (m). A single source [red dot in Fig. 7(a)] is placed at (4.5, 1.5, 1.5). The position of the microphone [blue circles in Fig. 7(a)] is uniformly shifted with 0.1 m interval along x and y directions in the whole room. Therefore, total number of training samples is $3380\ (= 69 \times 49 - 1)$ except a source-position. The heights of the sound source and microphones are set to 1.5 m. The reflection coefficients of six walls are set to 0.4091. From these settings, the acoustic RIR is easily obtained at each position of blue circle by an image model [6]. The amplitude of FFT of RIR is used as feature. All the other parameters for mapping algorithm are chosen from the analysis of previous two applications ($K = 200$, $P = 1$, the $\mathrm{maximum\ rank\ of\ RIR\ space} = 5$, the $\mathrm{number\ of\ cluster}\ = 50$, and the number of test frames is 676 made at random position in the room).

Fig. 7(b) shows the distance error for the monaural source localization experiment. Each curve is drawn by the average of Euclidean distances between the positions of targets and mapping outputs for all test samples. In all conditions, the proposed method gives us the best results. [7] shows the comparison between different sound source localization techniques, which are proposed in the literature during the last decade. In [7], the best performance reaches around 0.2 m. However, their simulation room size [3.4 × 5 (m)] is less than half of ours [7 × 5(m)]. Thus, we can argue that the source localization performance of the proposed method has a similar resolution with state-of-the-art techniques, even though our approach use only single microphone instead of microphone array.

V. Conclusion

This letter proposed a new nonlinear mapping method for one-to-many corresponding datasets. The potential of the proposed method was demonstrated via an S-curve to Swiss-roll mapping simulation. In addition, a lip-to-speech mapping and a monaural source localization experiments were performed as real applications. Because of the clustering scheme, the proposed and baseline methods did not destroy the entire data structure. Moreover, the mapping results of the proposed method were closer to the targets than the baseline. Even though the datasets had great complexity as well as different characteristic and intrinsic informations, the proposed approach still showed better performance than the other methods. The performance of the proposed method is robust to several parameters that are difficult to determine. These characteristics greatly enhance the efficiency of the proposed algorithm.

Acknowledgment

The authors wish to thank D. Lee and J. Hamm for their helpful support.

References

[1] L. K. Saul and S. T. Roweis, “Think globally, fit locally: Unsupervised learning of low dimensional manifold,” *J. Mach. Learn. Res.*, vol. 4, pp. 119–155, 2003.

[2] J. Tenenbaum, V. de Silva, and J. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, pp. 2319–2323, December 2000.

[3] Z. Zhang and H. Zha, “Principal manifolds and nonlinear dimensionality reduction via tangent space alignment,” *SIAM J. Sci. Comput.*, vol. 26, no. 1, pp. 313–338, 2004.

[4] J. Hamm, I. Ahn, and D. Lee, “Learning a manifold-constrained map between image sets: Applications to matching and pose estimation,” in *Proc. Computer Vision and Pattern Recognition*, 2006.

[5] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, “CUAVE: A new audio-visual database for multimodal human-computer interface research,” in *Proc. ICASSP*, 2002, pp. 2017–2020.

[6] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust Soc. Amer.*, vol. 65, no. 4, pp. 942–950, 1979.

[7] A. Brutti, M. Omologo, and P. Svaizer, “Comparison between different sound source localization techniques based on a real data collection,” in *Hands-Free Speech Communication and Microphone Arrays*, 2008.