2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Reducing Tongue Shape Dimensionality from Hundreds of Available Resources Using Autoencoder

Minghao Yang
[1]National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, China, 100190
[2]The Center for Excellence in Brain Science and Intelligent Technology of Chinese Academy of Sciences, Beijing China, 100190
mhyang@nlpr.ia.ac.cn

Dawei Zhang
[1]National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, China, 100190
100190
dawei.zhang@nlpr.ia.ac.cn

Jianhua Tao
[1]National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, China, 100190
[2]The Center for Excellence in Brain Science and Intelligent Technology of Chinese Academy of Sciences, Beijing China, 100190
[4]University of Chinese Academy of Sciences, China
jhtao@nlpr.ia.ac.cn

***Abstract*—In spite of various observation tools, tongue shapes are still scarce resource in reality. Autoencoder, a kind of deep neural networks (DNN), performs well on data reduction and pattern discovery. However, since autoencoder usually needs large scale data in training, challenges exist for traditional autoencoder to obtain tongues' motion patterns only from tens or hundreds of available tongue shapes. To overcome this problem, we propose a two-steps autoencoder, where we first construct a stacked denoising autoencoder (dAE) to learn the essential presentation of the tongue shapes from their possible deformations; then an additional autoencoder with small number of hidden units is added upon the previous stacked autoencoder, and used for dimensionality reduction. Experiments run on 240 vowels' tongue shapes obtained from Chinese speakers' pronunciation X-ray films, and the proposed model is compared with traditional dAE and the classical principal component analysis (PCA) on dimensionality reduction and reconstruction in details. Results validate the performance of the proposed tongue model.**

Keywords—vocal tract; tongue shape; PCA; neural network

I. INTRODUCTION

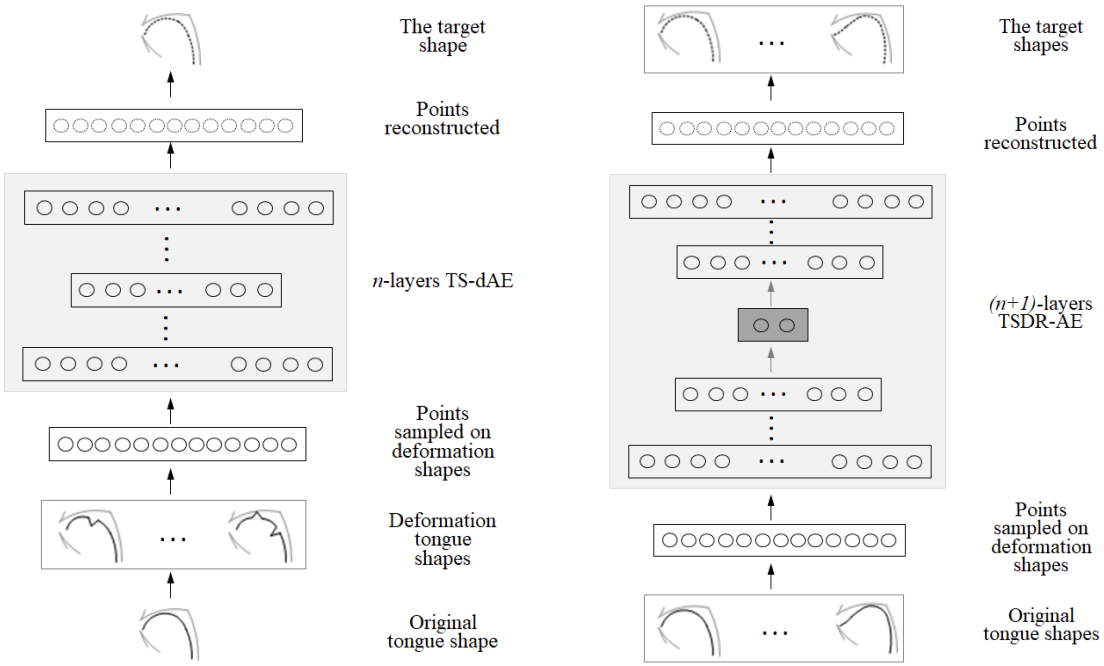
In spite of various techniques, it is still difficult to observe tongue contours directly as most parts of tongue are hidden inside the oral cavity. As a soft-tissue structure, tongue produces large deformation, which contributes to complex vocal tract configuration. As the tongue plays a crucial role in the transformation process from articulatory configuration to speech acoustics, it is important to model tongue shape and its configuration in speech production research.

In traditional experimental phonetics, various tongue models have been proposed to investigate the relationship between vocal tract configuration and speech production. Early in the 70 years of the last century, linguists and speech pathologists labeled tongue contours manually from X-ray films and used PCA to obtain the dominant patterns of tongue motion [1, 2]. It is reported that in vowel production, the first top two principal components explain over 90% variance, which means that vowel tongues' deformation could be described by the first two dimensionality parameters.

Parallel factor analysis (PARAFAC) is elegant tool for tongue shape analysis [3-7]. With PARAFAC on 13 cross sections for 10 English vowels, it was found that tongue movement could be described in terms of two factors, one generates a forward movement of tongue root accompanied by an upward of the front part of tongue, the other generates an upward and backward of whole tongue body [3]. However, PARAFAC did not consider tongue shape reconstruction from low-dimensional factors [7, 8].

There are some others tongue shape models, such as manifold representation of vowels [9], visualization of tongue trace [10, 11], speech driven tongue surface [12-14], and RBF based B-spline fitting [15]. These tongue models mainly focus on estimating global tongue movement trajectory from text, recorded speech or noised images, where tongue shape reconstruction and contributions of different tongue parts to speech production were not mentioned.

A high-performance tongue model includes two essential characteristics: low dimensionality and accuracy [5, 16-18]. Low dimensionality representation of tongue shape helps to reveal tongue motion patterns. Accurate reconstruction of tongue shape from low dimensionality parameters benefits to discover the mapping relationship between tongue motion patterns and articulatory configuration. In spite of various tongue models in history, researchers always try to achieve



(a) (b)

Fig. 1. Pipeline of the proposed two-steps autoencoder tongue model. (a) TS-dAE; (b) TSDR-AE.

tongue model with higher performance of dimensionality reduction and reconstruction ability.

It was reported that dAE performed well on data reduction and pattern discovery [19, 20]. However, as far as we known, there is still no work discussing whether DNN or autoencoder is suitable for modeling tongue shapes. One important reason for this situation is that autoencoder usually needs large scale data in training process, while tongue shapes are usually scarce resource in reality. We also want to know to what extent performance autoencoder would achieve if it could be used for dimensionality reduction of tongue shapes from a small amount of, i.e. only tens or hundreds of, tongue shapes. This paper aims to construct a high performance tongue model using autoencoder. And we are looking forward that autoencoder is efficient on dimensionality reduction of tongue shapes by comparing it with those of classical tongue models on shapes reduction and reconstruction.

DAE has been well reported for its high performance on data reduction [19, 21, 22], where corruption or dropout noise technique is used to enlarge the training data and contributes to the performance improvement from an amount of real data [19, 23, 24]. While in human pronunciation, the front parts of tongue movements are more dominant in vocal tract configurations than the latter parts of tongues. Therefore, the traditional corruption and dropout techniques may not fit tongue shapes denoising. To realize high performance of tongue shapes dimensionality reduction, we propose a two-steps autoencoder tongue model, where a large-scale noise shapes are first constructed from possible physiological deformation and is used to train an n-layer stacked autoencoder; then at the second step, a final autoencoder with small number of hidden units is added upon the previous stacked autoencoder, and is fine-tuned with real tongue shapes, which is similar to the idea of depth augmented network presented in article [25]. However, the proposed two-steps autoencoder differs with the traditional depth augmented network on that we aims to reconstruct tongue shapes using only a few available resources. We denote the first steps as Tongue Shapes Denoising Autoencoder (TS-dAE) and the second step as Tongue Shape Dimensionality Reduction Autoencoder (TSDR-AE) respectively. The overview of the proposed tongue model is presented in Fig. 1, where TS-dAE is given as Fig. 1(a) and TSDR-AE Fig. 1(b). The difference between TS-dAE and TSDR-AE is that noise shapes are used to train the n-layer TS-dAE, while real tongue shapes are used to fine-tune the (n+1)-layer TSDR-AE.

The remainder of the paper is organized as follows: Section II first outlines the framework of the proposed method, tongue shapes' normalization procedures and the design principles of networks; Section III gives the experiments and section IV concludes this paper.

II. Model Tongue Shape With Neurall Network

A. Tongue Shape Normalization

Similar to the presentation in [1, 3], shown as Fig. 2(a), vocal tract is characterized in terms of a set of reference lines shown in Fig. 2(b)(c)(d), before they are inputted to network. The cross-sections were divided by 18 grid lines in [3], where only the grid lines 4 to 17 (a total of 13 cross-sections) were used to describe the vocal tract configuration. The procedure to generate 13 reference grid lines in this work is as follows. Firstly, the tip point of upper teeth and palate are confirmed. Then the contour from teeth tip point along palate to epiglottis is taken as the reference background for different vocal tract configurations. Finally, normalized cross sectional distances

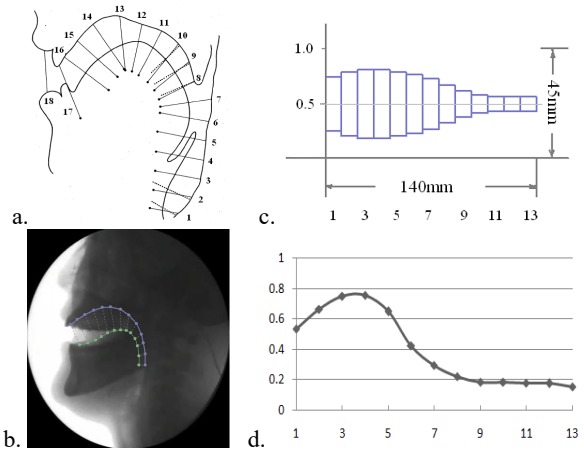


Figure 2: Tongue shape normalization. (a) the total 18 grid lines presented in [3], where the tongue shape contains the cross-section segments from grids 4 to 17; (b) the 13 cross-sections in the proposed model for the 32-th frame of /a/; (c) normalized cross-sections for (b) in vocal tract; (d) the normalized values of 13 units for (b) and (c) in 2D coordinate system.

between tongue surface and palate are taken as networks' input. For these 13 grid lines, the line segment lengths from reference background to tongue surface, which is orthotropic to background, were used for tongue factor analysis.

Being different from that 13 grid line lengths are directly used in PARAFAC [3], the normalized values of these grid lines, which belong to [0, 1], are used in network training. Eq.

$$\zeta_{ifj} = \Gamma_{ifj} / \max_{i,f,j}(J, \Gamma_{ifj}) \text{ and } \Gamma_{ifj} = (G_{ifj}/V_{ifj})\eta \qquad (1)$$

(1) presents the normalization procedure. In Eq. (1), i ($i \in [1,13]$) presents the i-th tongue line grid; f and j are the f-th frame of the j-th phoneme in tongue shape data set; $\mho$ is the widest cross-section distance of the vocal tract in mid-sagittal plane. In this work, $\mho$ is set to 45mm, which follows the configuration of $\mho$ in [3]. ζ_{ifj} and Γ_{ifj} are the normalized and un-normalized grid line length; η is real tongue length from tongue tip to tongue root. The whole tongue length of adult is about 175mm for male and 140mm for female [26]. G_{ifj} and V_{ifj} are length of grid line and tongue length in pixels, which can be obtained directly from X-ray films. Finally, the gridline length Γ_{ifj} and the corresponding normalized value ζ_{ifj} are obtained.

Fig. 2(b) shows length distribution of these 13 cross sections for the 32-th frame of phoneme /a/, which is similar to the set of reference lines introduced in [3] or in Fig. 2(a). Fig. 2(c) presents the tongue shape in tube model style from tongue tip to tongue root, which presents vocal tract configuration caused by tongue movement intuitively from a side view. The normalized lengths of 13 grid line $\zeta_{i,32,a}(i \in [1,13])$ from tongue tip to tongue root are listed in Fig. 2(d), which are the input of TSDR-AE in the proposed tongue model.

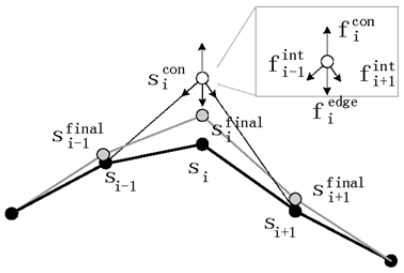


Fig. 3. An example of point deformation: the final position (in gray) caused by a random constraint force f_i^{con} (in white) appended to control point s_i at its original position (in black).

B. Tongue Shapes Deformation

At the first step of the proposed model, TS-dAE aims to generate high performance presentation of tongue shapes using denoising technique. To this end, a large scale noise shapes are constructed by adding noises to the original small real tongue shape set. As what we have introduced in the first paradigm of this section, traditional dAE use corruption or dropout noise technique to enlarge the training data [19, 23, 24], and these skills do not fit the vocal tract configuration in pronunciation. Therefore, we consider tongues' possible physiological deformations in TS-dAE training which are generated by active shape mode (ASM).

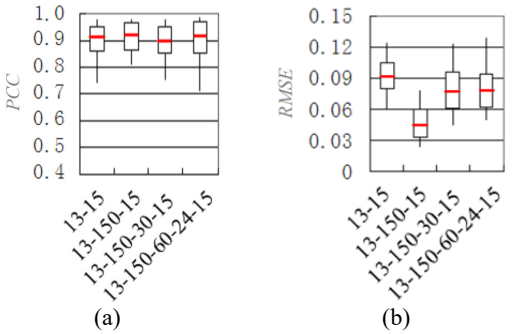
ASM is an efficient tool in simulating possible physiological deformation of tongue shapes [27, 28]. We use ASM to generate tongue shapes' deformations following the

$$E(S) = argmin \sum_{i=1}^{m} [E_{int}(s_i) + E_{edge}(s_i) + E_{con}(s_i)] \qquad (2)$$

description in [27, 28]. The deformations are described as Eq. (2). In Eq. (2), S means a tongue contour and s_i ($1 \le i \le m$) is the i-th control point on S, where m is total number of control points. For s_i, $E_{int}(s_i)$ presents the energy for internal forces from adjacent points f_{i-1}^{int}, f_{i+1}^{int}; $E_{edge}(s_i)$ is the edge force from original edge or contour f_i^{edge}, and $E_{con}(s_i)$ presents the constraint force f_i^{con} from user respectively. These forces are demonstrated in Fig. 3. In Fig. 3, the random external constraint force f_i^{con} at s_i contribute to s_i an initial deviation to s_i^{con}. And the final position of s_i at s_i^{final} is obtained by solving Eq. (2). References [27, 28] present the solution in details.

C. TS-dAE

The ASM based deformation shapes contribute to a large-scale noise shape data set, which is partly shown as the curves in solid line at middle bottom of Fig. 1(a). Based on these noise shapes, a tongue shapes denoising autoencoder is obtained. With its encoding and decoding procedure, the reconstructed shapes, showed as the curves in dotted line at right bottom of Fig. 1(a), are close to the original tongue shapes, namely the curves in solid line at left bottom of Fig. 1(a). The differences between the reconstructed tongue shapes and the original tongue shapes are requested as small as possible.



(a) (b)

Fig. 4. The average values, maximal values, minimum values and variance range of PCC and $RMSE$ for 13-15, 13-150-15, 13-150-30-15 and 13-150-60-24-15 stacked autoencoder with 5000 deformation shapes training. (a) PCC; (b) $RMSE$.

D. TSDR-AE

The output of TS-dAE is still of high dimensionality. A final autoencoder with small number of hidden units is further stacked at the top of TS-dAE, which contributes to the (n+1)-layer TSDR-AE. The training processing of TSDR-AE is shown as Fig. 1(b). The structure of the TSDR-AE is considered as following procedure. First, 13 segments are sampled at intervals from tongue tip (TT) to tongue root (TR). Then the normalized cross-sections distances of these 13 segments are taken as input units of TSDR-AE. Supposing that the number of visible and hidden units of the i-th layer autoencoder are ℓ_i^v and ℓ_i^h, where $\ell_i^h = \ell_{i+1}^v (1 \le i \le n)$ and $\ell_1^v = 13$, and the $(n+1)$-th layer autoencoder contains ℓ_{n+1}^v visible units and ℓ_{n+1}^h hidden units, where $\ell_{n+1}^h \ll \ell_n^h$ and $\ell_{n+1}^h \ll \ell_1^v$ (the symbol "$\ll$"means far less than), and the $(n+1)$ -th layer autoencoder realizes tongue shape dimensionality reduction. The reconstructed shapes, namely the dotted line curves at the right bottom of Fig. 1(b) are close to the original input shapes, the solid line curves at the left bottom of Fig. 1(b). Finally, a high dimensional tongue shape could be represented by the values of the units at the top layer of TSDR-AE: ℓ_{n+1}^h.

III. Experiments

A. Data Preparation

Despite radiation harm, X-ray remains an important technique in studying speech production for its higher time resolution [29-31]. In this study, the tongue shapes were taken from a pronunciation X-ray films spoken by Chinese females [32]. The video contained 20 phonemes (including mandarin vowels) and 181 syllables. The resolution for X-ray image is 640 × 480. The speakers' tongue shapes are normalized uniformly with Eq. (1). Each vowel lasts about 35 to 50 frames, and time for each frame is about 30ms. Five typical vowels (/a/, /i/, /u/, /e/, /o/) are taken into account to verify the proposed model in tongue shapes' dimensionality reduction and visualization. The middle part frames in pronunciation period present the essential and the steady pronunciation characteristics of vowels. 24 frames in the middle part frame are selected at intervals for each vowel. Finally, the picked tongue shapes (total 240 tongue shapes for 5 vowels) and their deformation shapes are used as training and test data.

Since large deformations usually happen at the font parts of tongue, the deformation units are most arranged at the previous sixth items in total 13 units. In total about 6000 deformation shapes, which is equal to $120 * \sum_{h=1}^{4} \left(\frac{h}{13}\right)$, are constructed from 120 real tongue shapes, among which 5000 shapes are used to train the neural networks, and the remaining 1000 shapes are used to evaluate TS-dAE. In TSDR-AE step, the previous 120 real tongue shapes are used to fine-tune TSDR-AE, and the remaining 120 tongue shapes are used for TSDR-AE evaluation.

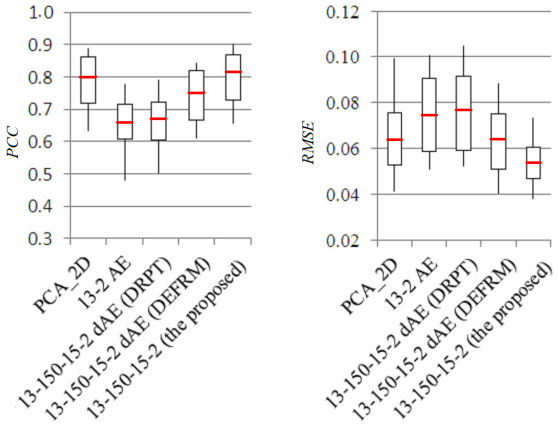
B. Structure of TS-dAE and TSDR-AE

It is important to achieve a good balance between the TS-dAE, TSDR-AE structure and their performances. We follow the training guidelines proposed in [24, 33] to optimize network structure.

Autoencoder is able to fit arbitrary data distribution in the case of enough units are provided in hidden layer [34]. For a practical reconstruction problem, the autoencoder usually archives good results when the number of hidden units is set to be about ten times of input units number [24]. As the input vector contains 13 units, and the final size of output layer of TSDR-AE is needed to be small, then the number of hidden units for the first layer autoencoder is restricted to 15 or 150 in this work. Then autoencoders with 10, 5 and 2.5 times decreasing for hidden units' number are used to construct multi-layer network step by step. In this way, TS-dAE is possibly constructed as 13-15, 13-150-15, 13-150-30-15 and 13-150-60-24-15 networks. Fig. 4 gives the Pearson Correlation Coefficient (PCC) and Root Mean Square Error ($RMSE$) measured by the original tongue shapes and the reconstructed tongue shapes obtained by 13-15, 13-150-15, 13-150-30-15, 13-150-60-24-15 networks respectively. And we can see from Fig. 4 that 13-150-15 obtains relatively highest performance of PCC and $RMSE$ among these networks. Then 13-150-15 network is used to construct the (n+1)-layer TSDR-AE.

To validate the performance and to benefit the comparison of the proposed model with the traditional models on two factors, we add an additional 15-2 autoencoder on the top of 13-150-15 TS-dAE and a stacked 13-150-15-2 TSDR-AE comes into being. We compare the proposed two-steps 13-150-15-2 autoencoder with PCA model in two factors, a standard two-layer (13-2) autoencoder, a standard 13-150-15-2 dAE with dropout skill and a 13-150-15-2 dAE using the proposed shapes deformation skill on tongue shapes reconstruction.

To facilitate the writing, we denote the two factors PCA model as PCA_2D; the standard 13-2 autoencoder as 13-2 AE; the standard 13-150-15-2 dAE with dropout skill as 13-150-15-2 dAE (DRPT), where "DRPT" mean dropout; and the 13-150-15-2 dAE using deformation denosing as 13-150-15-2 dAE (DEFRM), where "DEFRM" means deformation. The PCA_2D model, the 13-2 AE and the 13-150-15-2 dAE (DRPT) are trained on 120 real tongues with dropout technique. The 13-150-15-2 dAE (DEFRM) is trained with 5000 deformation tongues. The training procedure of TSDR-AE is similar to that of 13-150-15-2 dAE (DEFRM), except



(a) (b)

Fig. 5. The average values, maximal values, minimum values and variance range of ***PCC*** and ***RMSE*** for PCA_2D, 13-2 AE, 13-150-15-2 dAE using dropout skill, 13-150-15-2 dAE using deformation shapes and the proposed two-steps 13-150-15-2 model on 120 test tongue shapes. (a) ***PCC***; (b) ***RMSE***.

that TSDR-AE owns the second fine-tuning step using real tongue shapes.

Fig. 5 lists the average values, maximal values, minimum values and variance range of *PCC* and *RMSE* for these models. In Fig. 5(a), the proposed 13-150-15-2 network obtains similar *PCC* scores to that of PCA_2D and these two models achieve higher scores than other models. Fig. 5(b) shows that the proposed 13-150-15-2 network obtains obviously better performances on *RMSE* than PCA_2D and other models.

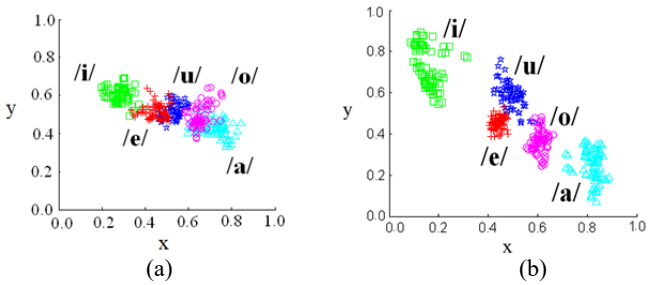
C. Visualization of Tongue Shapes in 2D Coordinate System

We map all 240 tongue shapes to 2D coordinate system using the proposed two-steps 13-150-15-2 autoencoder and PCA_2D model. Fig. 6(a) and Fig. 6(b) present all 240 points in 2D coordinate system obtained by PCA_2D and the proposed 13-150-15-2 network respectively. Points for different vowels are marked with different symbols, e.g., points in red plus correspond to the transformed tongue shapes of /e/ in 2D coordinate system.

In Fig. 6(a), most points of /i/ are separated from other points. This kind of distinguish is good for pronunciation labels. However, there are considerable number points of /u/ and /e/ which are overlapped together. And nearly half parts of /a/ are overlapped with points of /o/. It means that quit a few of tongue shapes obtained by PCA_2D, for instance, /a/ and /o/, /u/ and /e/ could not be distinguished in 2D coordinate. While in Fig. 6(b), there are obviously five clusters, and the points of /i/ and those of /a/ are clearly separated. Only small parts of /u/ are adjacent to the boundary points of /e/ and /o/. Fig. 6 demonstrates that the points obtained by the two-steps 13-150-15-2 network are better visualized and distinguished than those of PCA in 2D coordinate system.

D. Discussions

The deformation based 13-150-15-2 dAE (DEFRM) network is compared with 13-2 AE, 13-150-15-2 dAE (DRPT) on tongues' shapes reconstruction in Fig. 5. It shows that the deformation 13-150-15-2 dAE (DEFRM) obtains nearly 0.09



(a) (b)

Fig. 6. Visualization of the reduction points in 2D coordinate system obtained by (a) PCA_2D and (b) the proposed two steps 13-150-15-2 autoencoder.

increasing on *PCC*, and at the same time, nearly 0.013 decreasing on *RMSE* respectively than those of 13-2 AE and 13-150-15-2 dropout autoencoder. The experiments show that the proposed deformations denoising skill really increases the performance against the single layer autoencoder and traditional dropout autoencoder. It is because that the physiological mechanism based denoising contributes to a large-scale deformation data set. And it makes TS-dAE be robust to the small changes of input data, which are irrelevant to the large deformations at the font parts of tongue in speech pronunciations.

After a new additional autoencoder fine-tuning on real tongues shapes, the extra stacked autoencoder improves the lower bound of the probability approximation for input data, bringing higher reconstruction performance of the (n+1)-layer TSDR-AE than the traditional standard denosing autoencoder trained on deformation shapes. We can see from Fig. 5 that TSDR-AE obtains nearly 0.07 increasing on correlation coefficient, and nearly 0.01 decreasing on RMSE respectively than the deformation 13-150-15-2 dAE (DEFRM) network. It demonstrates that a new additional autoencoder continues to improve the performance from the standard denosing autoencoder trained on deformation shapes.

Fig.6 demonstrated that TSDR-AE outperformed PCA_2D on tongue shapes reduction and visualization in 2D coordinate system. The points obtained by TSDR-AE are better visualized and clustered than those of PCA in 2D coordinate system. The high reconstruction performance and good dimensionality reduction capacity of TSDR-AE ensure a more intuitive bi-directional mapping between tongue articulatory configurations and low-dimensionality parameter representations than PCA.

In general speaking, with the physiological deformations denosing skill and deep architecture, the proposed two-steps autoencoder contributes a novel tongued model, which own competitive performance against classical PCA tongue model. In spite of scarce resources of tongue shapes in reality, the proposed two-steps model obtains satisfactory performance of tongue shape reconstruction and dimensionality reduction.

IV. CONCLUSIONS

In this work, we discuss and confirm the possibility of using deep learning in vowels' tongue shapes reduction and visualization by a two-steps autoencoder. As far as we known, there is no work discussing whether DNN or autoencoder is suitable for modeling tongue shapes before. The main

contributions of the proposed two-steps autoencoder are: (1) it realizes a deep architecture for tongue shapes dimensionality reduction and reconstruction from scarce tongue resources. Because of large-scale deformation shapes denosing, the network is able to learn the essential representations which are robust to small irrelevant changes for tongue shapes. (2) it provides an bi-directional mapping mechanism between tongue shapes' articulatory configuration and its low dimensional parameters space. The proposed model is compared in details on tongue shape reconstruction performance with the traditional stacked autoencoder and PCA. Experiments indicate that the proposed model outperforms the traditional models mentioned above. The autoencoder construct process in this work could be widely and potentially used in the speech production research fields, such as visual/articulatory speech synthesis, computer-assisted pronunciation learning, etc.

ACKNOWLEDGEMENT

This work is supported by the National Key Research & Development Plan of China (No. 2017YFB1002804), the National Natural Science Foundation of China (NSFC) (NO.61332017, No.61425017).

REFERENCES

[1] M. Stone, M. H. G. Jr, and Y. Zhang, "Principal component analysis of cross sections of tongue shapes in vowel production," *Speech Communication,* vol. 22, pp. 173-184, Aug 1997.

[2] A. Hewer, I. Steiner, T. Bolkart, S. Wuhrer, and K. Richmond, "A statistical shape space model of the palate surface trained on 3D MRI scans of the vocal tract," in *18th International Congress of Phonetic Sciences*, Glasgow, United Kingdom, 2015.

[3] R. A. Harshman, L. Peter, and G. Louis, "Factor-Analysis of Tongue Shapes," *Journal of the Acoustical Society of America,* vol. 62, pp. 693-707, 1977.

[4] Y. Zheng, M. HasegawaJohnson, and S. Pizza, "Analysis of the three-dimensional tongue shape using a three-index factor analysis model," *Journal of the Acoustical Society of America,* vol. 113, pp. 478-486, Jan 2003.

[5] K. Iskarous, "Patterns of tongue movement," *Journal of Phonetics,* vol. 33, pp. 363-381, 2005.

[6] S. Maeda, "Articulatory Model of the Tongue Based on a Statistical-Analysis," *Journal of the Acoustical Society of America,* vol. 65, pp. S22-S22, 1979.

[7] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics and Intelligent Laboratory Systems,* vol. 38, pp. 149-171, Oct 1997.

[8] R. A. Harshman, "Foundations of the PARAFAC procedure Models and procedures for an explanatory multi-modal factor analysis," *UCLA Working Papers in Phonetics,* vol. 16, 1970.

[9] X. Lu and J. Dang, "Vowel Production Manifold: Intrinsic Factor Analysis of Vowel Articulation," *IEEE Transactions on Audio Speech and Language Processing,* vol. 18, pp. 1053-1062, Jul 2010.

[10] L. Wang, H. Chen, and S. Li, "Phoneme-level articulatory animation in pronunciation training," *Speech Communication,* vol. 54, pp. 845-856, Sep 2012.

[11] K. Xu, Y. Yang, A. Jaumard-Hakoun, P. Roussel, M. Stone, and B. Denby, "3D tongue motion visualization based on ultrasound image sequences," *Proceedings of Interspeech,* 2014.

[12] H. Li, M. Yang, and J. Tao, "Speaker-Independent Lips and Tongue Visualization of Vowels," in *2013 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp)*, Vancouver Canada, 2013, pp. 8106-8110.

[13] P. Birkholz, "Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis," *Plos One,* vol. 8, Apr 16 2013.

[14] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y. C. Kim, Y. H. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *Journal of the Acoustical Society of America,* vol. 136, pp. 1307-1311, Sep 2014.

[15] C. Qin, M. á. Carreira-Perpinán, K. Richmond, A. Wrench, and S. Renals, "Predicting tongue shapes from a few landmark locations," presented at the INTERSPEECH 2008, Conference of the International Speech Communication Association, Brisbane, Australia, 2008.

[16] D. A. Nix, G. Papcun, J. Hogden, and I. Zlokarnik, "Two cross-linguistic factors underlying tongue shapes for vowels," *Journal of the Acoustical Society of America,* vol. 99, pp. 3707-3717, Jun 1996.

[17] E. L. Saltzman and K. G. Munhall, "A Dynamical Approach to Gestural Patterning in Speech Production," *Ecological Psychology,* vol. 1, pp. 333-382, 1989.

[18] P. J. B. Jackson and V. D. Singampalli, "Statistical identification of articulation constraints in the production of speech," *Speech Communication,* vol. 51, pp. 695-710, 2009.

[19] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzago, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *Journal of Machine Learning Research,* vol. 11, pp. 3371-3408, 2010.

[20] D. Li and D. Yu, "Deep learning: Methods and applications," *Macrosoft Reearch One Microsofy Way Redmond, WA 98052,* 2014.

[21] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science,* vol. 313, pp. 504-507, Jul 28 2006.

[22] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," *Proceedings of the 25th international conference on Machine learning,* pp. 1096-1103, 2008.

[23] G. E. Hinton, S. Osinder, and T. YeeWhye, "A fast learning algorithm for deep belief nets," *Neural Computation,* vol. 18, pp. 1527-1554, Jul 2006.

[24] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," *Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science,* vol. 7700, pp. 437-478, 2012.

[25] Y. Wang, D. Ramanan, and M. Hebert, "Growing a Brain: Fine-Tuning by Increasing Model Capacity," presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu HI, USA, 2017.

[26] G. Fant, *Acoustic theory of speech production*. , Netherlands: Mouton, 2nd edition. 1970 (Translated into Russian, Nauka, Moskva, 1964): The Hague 1960.

[27] F. Howing, L. S. Dooley, and D. Wermser, "Tracking of non-rigid articulatory organs in X-ray image sequences," *Computerized Medical Imaging and Graphics,* vol. 23, pp. 59-67, Mar-Apr 1999.

[28] M. Kass, A. Witkin, and D. TerzoPoulos, "Snakes - Active Contour Models," *International Journal of Computer Vision,* vol. 1, pp. 321-331, 1987.

[29] D. M. Friederike Roers, Johan Sundberg, "Predicted singers' vocal fold lengths and voice classification-a study of x-ray morphological measures," *Journal of Voice Official Journal of the Voice Foundation,* vol. 23, pp. 408-413, 2009.

[30] R. Sock, F. Hirsch, and Y. Laprie, "An X-ray database, tools and procedures for the study of speech production," in *Proceedings of International Seminar on Speech Production*, 2011, pp. 41-48.

[31] F. Yang, "An articulatory model of standard Chinese using MRI and X-ray movie," *Journal of Chinese Linguistics,* vol. 43(1), pp. 269-294, 2015.

[32] M. Yang, J. Tao, and D. Zhang, "Extraction of Tongue Contour in X-Ray Videos," in *2013 Ieee International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver Canada, 2013, pp. 1094-1098.

[33] G. E. Hinton, "A practical guide to training restricted boltzmann machines," *Neural Networks: Tricks of the Trade (2nd ed.),* pp. 599-619 2012.

[34] N. Le Roux and Y. Bengio, "Representational power of restricted Boltzmann machines and deep belief networks," *Neural Computation,* vol. 20, pp. 1631-1649, Jun 2008.