ESTIMATE ARTICULATORY MRI SERIES FROM ACOUSTIC SIGNAL USING DEEP ARCHITECTURE

Hao Li, Jianhua Tao, Minghao Yang, Bin Liu

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
{hli, jhtao, mhyang, liubin}@ nlpr.ia.ac.cn

ABSTRACT

This paper presents our work on acoustic-to-articulatory inversion mapping, in which, the articulatory data is the MRI series for articulators on mid-sagittal plan. Deep architectures based on restricted Boltzmann machine (RBM) and linear regression are employed to construct the audio-visual mapping. We test two architectures to initialize the neural network: the bottom-up stacked RBM with top regression layer architecture and the one with extra Gaussian-Bernoulli RBM on the top of the former architecture. GMM-based mapping is used as baseline method. The MRI data from USC-TIMIT database is used for the training. The experimental results show that the deep regression network is an effective model to construct the mapping from acoustic speech signal to articulatory MRI series, and also indicate that it is a better strategy to initial the top layer as Gaussian-Bernoulli RBM to compress the MRI data before the liner regression.

Index Terms— acoustic-to-articulatory inversion, MRI, deep regression network, deep neural network

1. INTRODUCTION

The acoustic-to-articulatory inversion mapping is to predict the movements of human articulators from acoustic speech signal. The inversion mapping techniques are useful in several domains. In auto speech recognition system, the articulatory information can improve its performance [1]. In the study of talking head, it can be used to generate the motion of articulators such as lips and jaw [2]. Moreover, visual cues from articulatory movements can enhance speech perception [3]. The inversion mapping is an ill-posed problem for its one-to-many nature and high nonlinearity, which makes it a difficult task. Many machine learning methods have been applied to tackle this problem, such as the artificial neural networks [4], hidden Markov models [5, 6] and Gaussian mixture model (GMM)-based mapping [7, 8]. Deep architectures have also been used to obtain high accuracy in the inversion task, and achieve good result [9]. For the training of the mapping models, rich articulatory datasets are needed, which should contain quantitative articulatory position data along with recordings of acoustic data produced. Various methods are used to record a speaker's articulatory movements, including X-ray films [10, 11], magnetic resonance imaging (MRI) series [12], 3D motion capture and electromagnetic articulograph (EMA). In the prior works, EMA data is the most widely used data for it has high temporal resolution. However, the disadvantage of EMA is its low spatial resolution, and it is very hard to infer the vocal tract shape from separated positions of EMA sensors. On the contrary, MRI series for speech researches can provide dynamic information from the entire mid-sagittal plane of a speaker's vocal tract, or any other scan plane of interest. Mid-sagittal MRI captures not only the movement of lips, tongue and jaw, but also organs like the velum and glottal, which cannot be monitored with other techniques. MRI databases for continuous speech have been built by many groups in recent years, such as the mngu0 database [13] and USC-TIMIT database [12, 14]. Even though the sampling rates are lower than EMA or X-ray film, MRI is a unique source of dynamic information about vocal tract shaping.

In order to take advantages of articulatory MRI data and the good performance of deep architecture, we implement the experiments on acoustic-to-articulatory inversion mapping, in which, the articulatory data is the MRI series for articulators on mid-sagittal plan. This is a novel work in the study of the inversion mapping. Deep architectures based on restricted Boltzmann machine (RBM) [15] and linear regression is employed to construct the audio-visual mapping. We test two criterions for the unsupervised pre-training of the neural network: the stacked RBMs with top regression layer, and the stacked RBMs with regression layer in the second top layer. USC-TIMIT database is adopted in the experiments.

The rest of this paper is organized as follows: Section 2 describes the deep regression neural network architectures; Section 3 describes the data and its preprocessing procedural; Section 4 give the details of our experiments and the evaluation of the systems. We conclude this paper in Section 5.

2. DEEP ARCHITECTURES

We use deep belief networks for regression to construct the mapping between acoustic parameters and MRI data. The network architectures are adapted to solve the audio-visual multi-regression problem. Deep regression networks have been successfully applied in the acoustic-to-articulatory inversion mapping by Uria et al. [9], in which the articulatory data is EMA data. Considering the gray values of the MRI frames as output features, this method can also be used to build the mapping between acoustic features and MRI data.

2.1. Deep regression neural network

The first architecture (architecture I) of deep regression neural networks in shown in Figure 1 (I), in which the input data x is the acoustic feature vectors and the object y is the gray value vectors of MRI images, h_k denotes the k^{th} hidden layer. In architecture I, the first layer is Gaussian-Bernoulli RBM, the second and higher layer are Bernoulli-Bernoulli RBMs, the top layer is a linear regression layer. A RBM is an undirected graphical model formed by a visible

layer v and a hidden layer h. The states of the units in one layer are conditionally independent given the state of the units in the other layer. For Bernoulli-Bernoulli RBM, the conditional distributions have the following expression:

$$P(v_i = 1|h) = \text{sigm}(-b_i - W_{i.}h) \quad (1)$$

$$P(h_j = 1|v) = \text{sigm}(-c_j - W_{.j}^T v) \quad (2)$$

where $\text{sigm}(x) = 1/(1 + e^{-x})$ is the logistic sigmoid function, $W_{i.}$ corresponds to the i^{th} row and $W_{.j}$ corresponds to the j^{th} column of the weight matrix. b_i and c_j are the bias for visible and hidden units, respectively. A Gaussian-Bernoulli RBM (GB-RBM) [16] is a RBM whose visible units are real valued and follow a Gaussian probability distribution with diagonal covariance, while the hidden units are still binary valued and follow a Bernoulli distribution. The conditional distribution of v given h has the expression:

$$P(v_i|h) = N(-b_i - W_{i.}h, \sigma_i) \quad (3)$$

where $N(\mu, \sigma)$ denote Gaussian distribution with mean μ and standard deviation σ. σ_i is the standard deviation for visible unit i. For problems with continuous input features, binary visible unit is not an appropriate representation of the data, so that GB-RBM is used in the unsupervised pre-training of the first layer.

The most popular approximation for maximum likelihood training of RBM parameters is contrastive divergence (CD) learning [15], which is used in this paper. The training process for the whole network is to train each hidden layer treating the latest pre-trained hidden layer as the visible layer of a new RBM. The top layer is a linear regression layer, of which the input value is obtained by feeding the input data through all the layers trained previously, and the object is y. The regression layer can be initialized by solving the linear regression problem directly using the least square error solution. Finally, all of the parameters in the network are fine-tuned using back-propagation (BP) algorithm. The activation function of each hidden layer is sigmoid function.

2.2. Proposed deep architecture for inversion mapping

For problems with high dimension object like MRI images, dimension reduction is usually needed. Considering the dimension reduction for the object of linear regression, we can use GB-RBM to encode the MRI frames before linear regression. Since the MRI frames are the object of the deep regression neural networks. We add an additional GB-RBM layer between the output layer and the linear regression layer. In this architecture the top hidden layer is the compressed representation for the MRI frames. Figure 1 (II) illustrates this architecture (architecture II). Its training process begins by pre-training the RBMs below and above the regression layer respectively, and then initials the regression layer parameters with the value of two hidden layers. When initial the regression layer, the object y_r is not binary value but real value which is the linear combination of y, where the weights and bias are from top GB-RBM. Finally, all of the parameters in the network are fine-tuned using BP algorithm.

There is a difference between the top RBM layer and the bottom RBM layer for this specific problem. For the bottom layer, the input data is acoustic spectrum and energy, which are approximately follow Gaussian probability distributions, therefore, after the z-score normalization of the spectrum features, the distribution of each input data component can be modeled by a standard Gaussian distribution $N(0, 1)$ and thus the conditional distribution of v given h is:

$$P(v_j|h) = N(-b_j - W_{j.}h, 1) \quad (4)$$

For the pre-training of the top layer, the training MRI data will be feed to the visible layer of RBM. In the MRI frames, the variation of gray value of some pixels are very small during the speech, such as the nose region and the background region (see Figure 2). Therefore, the gray values will only normalized by subtracting mean but will not divided by standard deviation to avoid small-value division. For each component of the MRI data (gray value of MRI frames), we calculate the standard deviation through all training data, and model its distribution with a Gaussian which have zero mean and its true deviation. The components are independent to each other. The conditional distribution of v given h is:

$$P(v_i|h) = N(-b_i - W_{i.}h, \sigma_i^y) \quad (5)$$

where σ_i^y is the standard deviation of the i^{th} component of y. v_i is y_i, the i^{th} component of y. The hidden units h are binary.

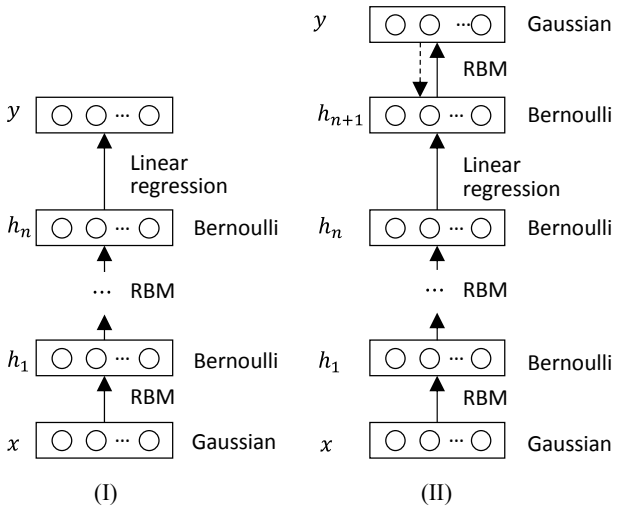
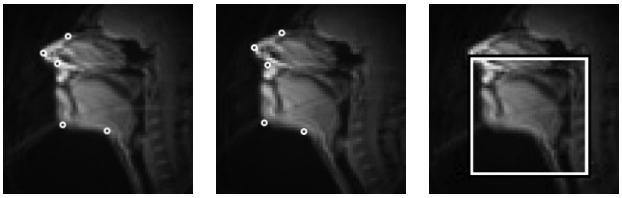


Figure 1. Deep architectures for mapping from acoustic features to MRI images: (I) a deep regression network with n hidden layer and top linear regression layer; (II) a deep regression network with extra RBM above the linear regression layer.

3. DATA

We use the MRI data form USC-TIMIT [14] database, which contains large-scale data of synchronized audio and MRI data for speech research. Subjects' vocal tracts were imaged in the mid-sagittal plane while reading 460 TIMIT sentences. The MRI image resolution in the mid-sagittal plane was 68×68 pixels (2.9×2.9mm). The image data were reconstructed as 23.18 frames/second. The audio was simultaneously recorded at a sampling frequency of 20 kHz inside the MRI scanner while subjects were imaged.

We use only one speaker's data from this database. The position of speaker's head may slightly change across different files in the database. The inconsistence of the head position may cause significant shake move of head in the estimated MRI series because the same phoneme may uttered from different head positions. To cope with this problem, we align the head position across files using the following method: firstly, we calculate a mean image for each MRI file, then manually mark five landmarks on the outline of the head in the mean images. Rigid transformations involve only translation and rotation were calculated to align the landmarks to a standard position (e.g. the positions in the first file), then we implement this transformation to all frames of the file. The mean images for tow files and their landmarks are shown in Figure 2 (a) and (b). This alignment procedure can alleviate the head shake in the estimated MRI series.



(a) (b) (c)

Figure 2. (a) and (b) show the mean images and five marker's positions (white circles) for align the head position of two MRI series files respectively. (c) is the interested region in the MRI images (within the white rectangle)

In the inversion task we only interested in the motion of articulators. Many parts of the MRI images are barely move during the speech such as the nose, part of the neck and part of the nasal cavity, as well as the black background. Therefore, we cut the border of the MRI images, and use an interest region with 40×40 pixels, which contains all the articulators who will move during the speech (lips, jaw, tongue, teeth, velum, glottis, etc.). The upper-left point of the interest region is (16, 20) in the original images. The interest region is shown in Figure 2 (c). The pixels outside the region were not used. Therefore, the dimension of one MRI sample vector is 1600.

The audio waveforms were divided into frames, the frame length and shift were 25ms and 4.31ms, respectively. The acoustic data were parameterized with 24 order line spectral pairs (LSPs) and log energy. We use a context window of 10 acoustic frames, thus, each input window will span a period of 43.1ms, which is the same to that of one MRI frame. The dimension of each input vector is 250.

4. EXPERIMENTS

We use the data of speaker f1 in the experiments. The data was separated into 3 groups: training, validation and test, which contain 360, 47 and 50 sentences, respectively (the audio of 3 sentences were missing in the original dataset). To measure the accuracy of our systems we use the average over test data of the root mean-squared error (RMSE) $\sqrt{(1/n)\sum_i(\hat{y}_i - y_i)^2}$, where $\hat{y}_i$ is the estimated gray value vector and y_i is the actual one of the MRI frame at time i, n is the number of test samples.

4.1. Network training configuration

The training configuration parameters for pre-training of RBM using CD learning and the fine-tuning using BP are shown in Table 1. They are the results of manually tuning to obtain a low reconstruction error. For the pre-training of linear regression layer, we use normal equation with regularization term, which has the expression:

$$\begin{bmatrix} b_r \\ W_r \end{bmatrix} = \left(X_r^{*T} X_r^* + \lambda \begin{bmatrix} 0 & \\ & I_{n\times n} \end{bmatrix} \right)^{-1} X_r^{*T} Y_r \qquad (6)$$

where W_r is the weight for the linear regression layer and b_r is the bias vector. $X_r^* = [1 \quad X_r]$, X_r and Y_r are the input and object of the regression layer respectively, they are obtained by feed the training data through the pre-trained RBMs. Y_r is the training target in architecture I. λ is the regularization parameter, which is used to control the scale of weights, I is identity matrix. We set $\lambda = 10$ and 35 in architecture I and II, respectively, which are the results of manually tuning, to achieve a lower regression error. The linear regression layer was trained before the fine-tuning of the whole network.

For architecture I, different number of layers and units were trained, and their performances on the validation set using the RMSE criterion are shown in Figure 3. We can observe that the best results are obtained with 2 hidden layers and 512 units per hidden layer, which has an average RMSE of 16.8.

Table 1. The configure parameters for the training of networks

	RBM	BP
Learning rate	0.0001	0.01
Momentum	0.5 (10 first epochs) 0.9 (rest of epochs)	-
Total epochs	200	100
Minibatch size	100	100
Initial weights	$N(0, 0.01)$	-
Initial visible bias	0	-
Initial hidden bias	0	-
Learning rate scaling	1	0.99

For architecture II, we froze the network under the regression layer to be two hidden layer with 512 unit per layer, and vary the unit number of the top hidden layer. Their performances on the validation set are shown in Figure 4. We can observe that when the top layer has 128 units, the system has the best result, which is an average RMSE of 16.76. Some of the weight of this layer after CD training is shown in Figure 5, from which we can observe that this layer is focusing on the pixies that have significant variations and thus will reflect the motion of the articulators. According to these experimental results, we set the final architecture of the proposed method as having four hidden layers and 512 units for the first two layer, 128 units for the top layer.

4.2. Baseline

We also train the inversion mapping with a widely used regression method, GMM-based mapping using minimum mean square error criterion, which have been used for acoustic-to-articulatory inversion mapping [7]. PCA whiten is applied on both the acoustic spectrum and MRI data vectors to reduce the dimensions. The acoustic and MRI data use the same number of components, which is 64 in the experiments. The RMSE performance on validation data of the GMM-based mapping is 17.89, 17.87 and 17.53 with the Gaussian mixture number of 32, 64 and 128 respectively. The GMM configuration that lead to the best mapping performance will be regarded as baseline.

4.3. Results and discussion

The performances of the two deep architectures and GMM-based mapping on the test set are shown in Table 2. The best result of deep architecture is an average RMSE of 17.74, which is significant lower than that of the GMM-based mapping. The distribution of the RMSEs for pixels in the MRI frame is shown in Figure 6. We can observe that the dark border of tongue in GMM-based mapping is slightly broader than that in the proposed method, which indicates that the estimated tongue shape of proposed method is more accurate.

From Figure 6 we can also observe that there is relatively higher error in the outline of nose, which is believed to be caused by the inconstancy of head position in the database, because nose will not move along with speech. From this phenomenon we can infer that the error on other part of the images is also partly caused by this reason.

Figure 7 shows four continues MRI images estimated by the proposed method and the actual MRI images. The estimated images have a blur effect compared with the actual ones. The contrast ratio

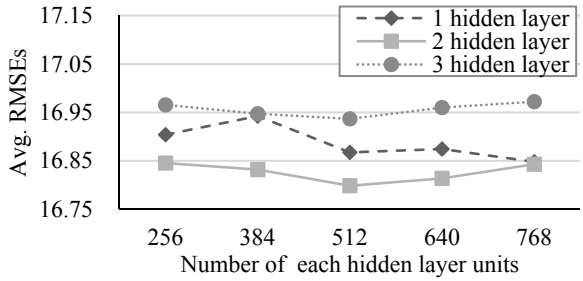


Figure 3. RMSE performances of architecture I as functions of the number of hidden layer units. The hidden layers have the same units' number.

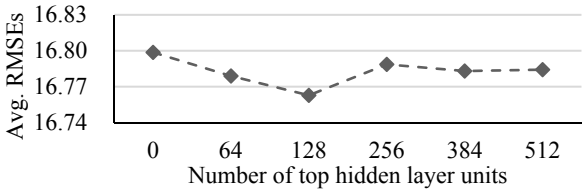


Figure 4. RMSE performance of architecture II as a function of the number of the top RBM layer. The networks structure blow the regression layer are set to 2 layers with 512 units each layer. The "0" in horizontal axis correspond to the best result achieved by architecture I.

Table 2. The average RMSEs performance on the test set.

	GMM	Architecture I	Architecture II
Avg. RMSE	18.48	17.76	17.74

of the estimated images is lower than actual ones and the contours of the organs are less clear. We can observe that the movement of articulators are accurate. The velum, glottal are at the right state for the current pronunciation. The shapes of lips and tongue are also very close with actual ones, and specifically, the collision between tongue tip and upper teeth is well recovered. Beside the tongue tip and upper teeth, most articulator collisions can be estimated from our observation, such as the tongue dorsum and the velum, tongue body and the hard palate.

The deep architecture II has better performance on validation data compared with architecture I. In Figure 3, the best RMSE performance for architecture I with 3 hidden layer is 16.94, which is achieved by using 512 units per layer. Using the same layer number and unit number, the architecture II achieved an average RMSE of 16.78 as shown in Figure 4. This results indicate it is a better strategy for this problem that pre-training the top layer as RBM and use it as an encoder for the object of the liner regression. When using architecture II, the best units' number for the top hidden layer is 128, which is less than that for acoustic spectrum parameters. This indicates that the mid-sagittal plan MRI image has a lower degree of freedom, and it contains less information than acoustic data, which is true because the image only contains 2-dimensional information of articulation. When compressing the MRI images into 64 dimensions, the GMM-based mapping has 17.53 average RMSE on validation data while that of the deep regression network with 64 units in top hidden layer is 16.78. From the experimental results we can infer that proposed deep regression network is an effective way to construct the audio-to-visual mapping for the estimation of the MRI series. The top GB-RBM plays an important role to abstract the MRI information. The architectures with more RBM layer between the

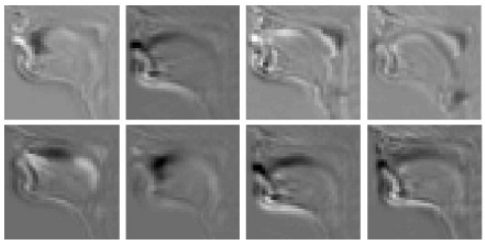


Figure 5. Visualization of some of the weight in the top GB-RBM after CD training, the number of hidden units is 512.

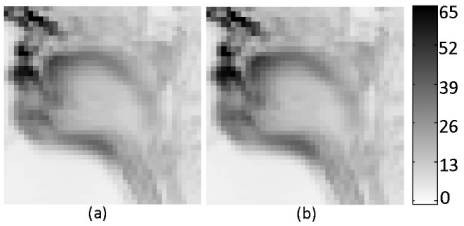


Figure 6. Distribution of average RMSEs for pixels in MRI frames for (a) deep regression network (b) GMM-based mapping.

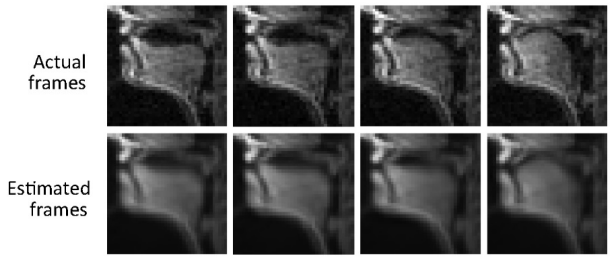


Figure 7. A series of estimated MRI images by deep regression network using architecture II and the actual images.

object and the regression layer is not presented in this paper, because they did not give better result than use only one RBM for the MRI data.

5. CONCLUSIONS

We present the novel work on estimating the articulatory MRI series from acoustic speech signal. We found deep architectures are able to obtain better inversion accuracy than GMM-based method. We implemented two architectures to pre-training the neural networks: the stacked RBM with top linear regression layer and the one with linear regression in the second top layer. The second architecture was proven to be a better one. In our future work, analysis based on the estimated MRI series using image processing technology will be carried on, such as contour extraction for vocal tract. We will also improve the evaluation metric for the inversion mapping, the vocal tract shapes will be compared systematically with that in actual MRI series, so that we can know more details about the visual information provided by the inversion mapping algorithm.

6. ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61273288, No.61233009, and No. 61425017), and the Major Program for the National Social Science Fund of China (13&ZD189).

7. REFERENCES

[1] S. King, *et al.*, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America,* vol. 121, pp. 723-742, 2007.

[2] P. Badin, F. Elisei, G. Bailly, and Y. Tarabalka, "An Audiovisual Talking Head for Augmented Speech Generation: Models and Animations Based on a Real Speaker's Articulatory Data," in *Articulated Motion and Deformable Objects*. vol. 5098, ed: Springer Berlin Heidelberg, 2008, pp. 132-143.

[3] P. Badin, Y. Tarabalka, F. Elisei, and G. Bailly, "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Communication,* vol. 52, pp. 493-503, 2010.

[4] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Proc. ICSLP*, Pittsburgh,USA, pp. 577–580, 2006.

[5] Z.-H. Ling, K. Richmond, and J. Yamagishi, "An Analysis of HMM-based prediction of articulatory movements," *Speech Communication,* vol. 52, pp. 834-846, 2010.

[6] S. Hiroya and M. Honda, "Determination of articulatory movements from speech acoustics using an HMM-based speech production model," in *Proc. ICASSP*, Orlando, U.S.A, pp. 437-440, 2002.

[7] T. Toda, A. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with Gaussian mixture model," in *Proc. ICSLP*, Jeju, Korea, pp. 1129–1132, 2004.

[8] I. Y. Ozbek, M. Hasegawa-Johnson, and M. Demirekler, "Estimation of Articulatory Trajectories Based on Gaussian Mixture Model (GMM) With Audio-Visual Information Fusion and Dynamic Kalman Smoothing," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 1180-1195, 2011.

[9] B. Uria, I. Murray, S. Renals, and K. Richmond, "Deep Architectures for Articulatory Inversion," in *Proc. INTERSPEECH*, 2012.

[10] J. Fontecave Jallon and F. Berthommier, "A semi-automatic method for extracting vocal tract movements from X-ray films," *Speech Communication,* vol. 51, pp. 97-115, 2009.

[11] Y. Minghao, T. Jianhua, and Z. Dawei, "Extraction of tongue contour in X-ray videos," in *Proc. ICASSP 2013*, pp. 1094-1098, 2013.

[12] S. Narayanan, *et al.*, "A Multimodal Real-Time MRI Articulatory Corpus for Speech Research," in *Proc. Interspeech 2011*, pp. 837-840, 2011.

[13] I. Steiner, K. Richmond, I. Marshall, and C. D. Gray, "The magnetic resonance imaging subset of the mngu0 articulatory corpus," *The Journal of the Acoustical Society of America,* vol. 131, pp. EL106-EL111, 2012.

[14] S. Narayanan, *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *The Journal of the Acoustical Society of America,* vol. 136, pp. 1307-1311, 2014.

[15] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation,* vol. 14, pp. 1771-1800, 2002.

[16] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems,* vol. 19, p. 153, 2007.