J Inf Process Syst, Vol.15, No.1, pp.91~99, February 2019
https://doi.org/10.3745/JIPS.04.0101

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Feature Extraction Based on DBN-SVM for Tone Recognition

Hao Chao*, Cheng Song*, Bao-Yun Lu*, and Yong-Li Liu*

Abstract

An innovative tone modeling framework based on deep neural networks in tone recognition was proposed in this paper. In the framework, both the prosodic features and the articulatory features were firstly extracted as the raw input data. Then, a 5-layer-deep deep belief network was presented to obtain high-level tone features. Finally, support vector machine was trained to recognize tones. The 863-data corpus had been applied in experiments, and the results show that the proposed method helped improve the recognition accuracy significantly for all tone patterns. Meanwhile, the average tone recognition rate reached 83.03%, which is 8.61% higher than that of the original method.

Keywords

Deep Belief Networks, Deep Learning, Feature Fusion, Support Vector Machine, Tone Feature Extraction

1. Introduction

Tone is one of the three key elements of Chinese syllable. To distinguish the meaning of syllables and words, besides the initial consonants and finals, the tone cues are also needed. In addition, tone is the only element to distinguish the characters which correspond to a same syllable. Without considering the element of tone, the number of Chinese syllables can decrease from over 1,300 down to about 400. In this way, the range of search space can be dramatically narrowed. Therefore, accurate recognition of tones is crucial for effectively improving the performance of the speech recognition system [1].

Feature extraction plays a key role in pattern recognition. Fundamental frequency (F0) is often used in tone modeling because pitch contour contains salient information regarding four normal tones. Different from normal tones, the neutral tones have no pitch contours but shorter duration and lower energy. Therefore, both the duration and energy features are also used for tone model [2,3]. In addition, it has been demonstrated previously in our work that the articulatory features also contain salient information regarding the four lexical tones [4]. Whenever the prosodic and pronunciation features are used together, the results of tone recognition are better than those that use the prosodic features alone. This indicates that these two types of features are complementary to each other. However, prosodic features and articulatory features are simply connected and lacking of more effective method for feature fusion.

Manuscript received May 16, 2017; first revision August 29, 2017; accepted October 23, 2017.

Corresponding Author: Cheng Song (chaozh2015@163.com)

* School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China ({chaohao1981, chaozh2015}@163.com, {Lubaoyun, liuyongli}@hpu.edu.cn)

In 2006, Hinton and Salakhutdinov [5] proposed the concept of deep learning and the deep belief network (DBN). Since then, deep learning has achieved brilliant results in mage, speech recognition and other fields [6–9]. As an emerging method in the field of machine learning and data mining, one of the important functions of the deep neural networks (DNNs) is to extract high-level abstract features on big sample data layer by layer.

Compared with shallow models such as support vector machine (SVM) and multilayer perceptron (MLP), as for the speech signal process, DNN is mainly featured in the advantages as follow. Firstly, DNN is quite capable in extracting features from large amount of data automatically, which reduces the reliance on expert experience and signal processing techniques. Secondly, DNN can well describe the complex mapping between the signal and the target class, and is quite applicable for the analysis on the diversity, nonlinear, high dimensional data in the big data background. Thirdly, deep models can better integrate different types of input features which represent speech signals from different aspects, thereby showing more powerful capability in feature learning than the method of traditional feature fusion.

So far, few researches have applied deep neural network to Mandarin tone recognition. Therefore, applying DBN to Mandarin tone recognition was presented in this paper. According to two types of raw input data (the prosodic features and the articulatory features), the DBN was applied to define the high-level abstract representation of the two types of raw input data, during which, the prosodic features and the articulatory features are fused together. In addition, it incorporates the tone features of adjacent syllables to cope with the co-articulation effect, and uses SVM classifier to recognize the tone. The traditional tone modeling methods were compared in our study with this method, based on which, a conclusion was made: the tone recognition rate reached 83.03%, which is 8.61% higher than that of the original algorithm.

As a follow-up study to our previous research described in [4], an improved tone modeling which applies DBN for perform effective fusion of two types of raw features was proposed in this paper. In addition, DBN-SVM framework was proposed for tone recognition as well in this paper.

The rest of the paper is organized as follows. Section 2 introduces the two types of raw feature. Section 3 introduces DBN and the relevant model, restricted Boltzmann machine (RBM). Section 4 describes experimental results and analysis. Finally, Section 5 briefly concludes the work.

2. Prosodic Features and Articulatory Features

2.1 Prosodic Features

Prosodic features contain pitch contour, duration and energy. In view of co-articulation, some certain features come from adjacent syllables. So, a number of prosodic features are extracted, as shown in Table 1. The prosodic features of each syllable can be described as follows:

- Pitch features: a second-order polynomial function is applied to fit pitch contour. The parameters of the function are treated as pitch contour features. In addition, each syllable is evenly divided into three segments. For each segment, the mean and rake ratio of its pitch contour are calculated.
- Energy features: the means of the log-energy and its first derivative.
- Duration features: duration, duration ratios of current syllables to the two neighboring syllables.
- Prosodic features of preceding syllable and following syllable: the above three kinds of features of preceding syllable and following syllable.

Table 1. Prosodic features used for tone modeling

Prosodic feature	Feature
Pitch contour features	9
Energy-related features	2
Duration-related features	3
The same three kinds of features from the preceding and following syllables	28

2.2 Articulatory Features

Fifteen-dimensional articulatory features had been proposed for tone modeling in our previous work. Here in this paper, the articulatory features which describe the pronunciation information of speech signal are also applied as the raw input features of deep belief networks. As for the specific extraction process of articulatory features, please refer to reference [4].

3. Deep Belief Network

As can be seen from Fig. 1, some RBM models form a DBN. In DBN, the output of the former RBM acts as the input of the latter RBM. Through this way, each hidden layer can obtain higher level representation of raw data than the previous one. Finally, the features extracted by DBN can be employed to train a supervised learning method such as SVM.

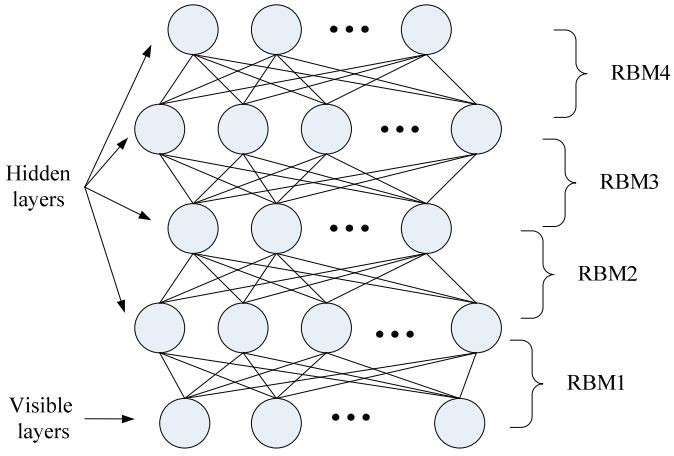

Fig. 1. Structure diagram of deep belief network.

3.1 Bernoulli RBM and Gaussian-Bernoulli RBM

Fig. 2 shows the structure of RBM, from which, it can be seen that each RBM contains two layers, respectively. One is visible layer, and the other is hidden layer. Each layer consists of several units (neurons). Similar to multi-layer perceptron, neurons in adjacent layers are interconnected, with no link between neurons existing in the same layer. Each RBM is exploited to represent the distribution of its

input. In this paper, the first RBM in DBN is Gaussian-Bernoulli RBM because both the prosodic features and the articulatory features are real-valued. And all other RBMs in DBN are Bernoulli RBMs.

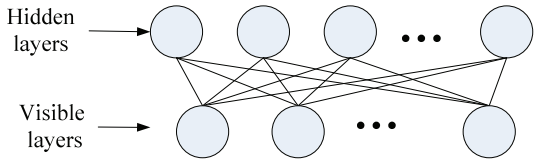

Fig. 2. Structure diagram of restricted Boltzmann machine.

In Bernoulli RBM, neurons of visible and hidden layers are binary, and the joint probability distribution function (JPDF) is described as

$$P(v,h)=\frac{1}{Z}\exp(-E(v,h)) \tag{1}$$

where $P(v, h)$ is the JPDF, v represents the visible neurons and h represents the hidden neurons. Z represents the normalization term. $E(v, h)$ represents the energy function. Among the above, Z is defined as

$$Z=\sum_{D}\sum_{K}\exp(-E(v,h)) \tag{2}$$

And the energy function is described as

$$E(v,h)=-h^{T}Wv-b^{T}v-c^{T}h=-\sum_{i=1}^{D}\sum_{j=1}^{K}W_{ij}v_{i}h_{j}-\sum_{i=1}^{D}b_{i}v_{i}-\sum_{j=1}^{K}c_{j}h_{j} \tag{3}$$

where W represents the connection weight coefficient matrix, D and K present unit numbers of visible and hidden layers. b represents the offset of visible neurons, and c represents the offset of hidden neurons.

As can be seen from $E(v, h)$, the hidden neurons are independent to each other when giving out the visible unit state, and the converse is also true. It means that each neuron in the visual layer is independent to the remaining neurons giving the state value of the hidden layer. According to the definition of conditional probability distribution, the binary state of each hidden unit can be set to 1, which is shown as follows:

$$p(h_{j}=1\,|\,v)=sigm(\sum_{i}W_{ij}v_{i}+c_{j}) \tag{4}$$

where $sigm()$ represents the sigmoid function. Similarly, the binary state of each visible unit can be set to 1, which is shown as follows:

$$p(v_{i}=1\,|\,h)=sigm(\sum_{j}W_{ij}h_{j}+b_{i}) \tag{5}$$

In Gaussian-Bernoulli RBM, neurons in hidden layer obey Bernoulli distribution. Neurons in visible

layer neurons obey Gaussian distribution. Then, $E(v, h)$ can be modified as

$$E(v,h)=\sum_{i=1}^{D}\frac{(v_i-b_i)^2}{2\sigma_i^2}-\sum_{j=1}^{K}c_j h_j-\sum_{i,j}\frac{v_i}{\sigma_i}W_{ij}h_j \tag{6}$$

where σ_i is standard deviation of visible layer units. Assuming σ_i=1, formulas (4) and (5) are modified as

$$p(h_j=1\,|\,v)=sigm(\sum_{i=1}^{D}W_{ij}v_i+c_j) \tag{7}$$

$$p(v_i=x\,|\,h)=N(\sum_{j=1}^{K}W_{ij}h_j+b_i,1) \tag{8}$$

where $N()$ is a function of Gaussian distribution.

3.2 Training of RBMs Model and DBN Model

According to the training data, the parameters set (W, b, c) of RBM can be acquired by exploiting the criteria of empirical negative log-likelihood minimization. Because it is very difficult to calculate the accurate gradient, an approximate algorithm is often used in RBM training to update parameters. This algorithm is contrastive divergence (CD) introduced in [10]. Finally, updating formula of the parameters set (W, b, c) is achieved

$$W^k=W^{k-1}+\varepsilon(<vh^T>_{data}-<vh^T>_{model}) \tag{9}$$

$$b^k=b^{k-1}+\varepsilon(<v>_{data}-<v>_{model}) \tag{10}$$

$$c^k=c^{k-1}+\varepsilon(<h>_{data}-<h>_{model}) \tag{11}$$

where $<\bullet>_{data}$ is the expectation of training data observation, $<\bullet>_{model}$ is the expectation under distribution defined by model, and ε is the learning rate.

DBN training is divided into two stages: pre-training and fine-tuning. As for the specific process of training, please refer to [7,8].

4. Experimental Results and Analysis

4.1 Speech Corpora

The 863 dataset adopted in this paper is divided into two parts: the training set and the test set, among which, the training set contains 43,573 sentences and is recorded from 83 male volunteers in a period of 49.4 hours, and the test set is consisted of 4,800 sentences, recorded from 24 male volunteers in a period of 347 minutes. There are 526,165 syllables in the training set and 60,816 syllables in the test set in total. Four normal tones of these syllables are marked as 1, 2, 3, 4. Neutral tone is marked as 0. Table 2 shows the tone distribution of syllables in training set and test set.

Table 2. The tone distribution of syllables in training set and test set

	Total	Tone				
		0	1	2	3	4
Training set	526,165	51,516	99,693	119,160	86,234	169,562
Test set	608,16	5,943	11,508	13,776	9,975	19,614

The data corpus is recorded in the laboratory environment, and is stored in the 16 kHz sampling rate and 16-bit resolution.

4.2 Experimental Setup

In this paper, three different tone modeling methods based on SVM are adopted with their respective performance evaluated on the task of tone recognition for test set. Firstly, prosodic features described in Table 1 are employed to build a SVM as tone classifier (P-SVM). Secondly, prosodic features and articulatory features are both used for the training of a SVM (PA-SVM), which has been described in our previous work [4]. Finally, prosodic features and articulatory features are fed as raw features into a 5-layer-DBN which can extract high-level tone features. Then, the high-level tone features are fed into a nonlinear support vector machine (DBN-SVM). All the SVMs adopt RBF kernel and are trained by the LibSVM [11]. The value of punishment coefficient C is 16 and the value of gamma is 0.2. The process of all experiments is shown in Fig. 3. By using a single classifier, direct comparisons can be carried out.

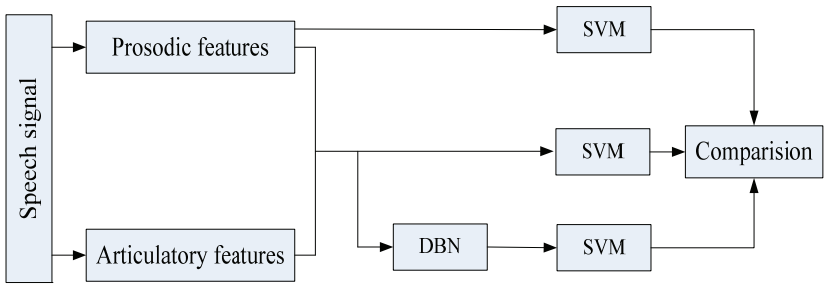

Fig. 3. The process of experiment.

For each syllable, prosodic features (42-dimensions) and articulatory features (19-dimensions) are firstly extracted as the DBN inputs. The DBN used in the experiment includes 5 hidden layers. The number of nodes in each layer is described as 61-60-50-50-50. Gaussian-Bernoulli RBM is adopted in the first layer while Bernoulli RBM is adopted by other layers. The weights of the first layer is set by sampling from a Gaussian distribution N (0,0.04). The offset of visible neurons are set as 0. For pre-training, the unsupervised layers ran for 50 epochs in 0.001 learning rate. And as for the fine-tuning, the supervised layers ran for 200 epochs in 0.01 learning rate.

4.3 Analysis on Experimental Results

Tables 3–5 show the performance of P-SVM, PA-SVM, and DBN-SVM, respectively. Compared with the tone recognition results of P-SVM in Table 3, the recognition accuracies of all five tones increase

significantly when both the prosodic features and the articulatory features are applied together. This means that the prosodic features and the articulatory features describe tones from different aspects and the two kinds of features should be complementary in tone recognition.

Table 3. The results of the tone recognition rate based on P-SVM by five types

	Tone recognition rate (%)				
	Type 0	**Type 1**	**Type 2**	**Type 3**	**Type 4**
Type 0	75.62	5.54	6.75	4.53	7.56
Type 1	1.19	78.99	4.77	5.51	9.54
Type 2	0.75	7.01	67.21	12.30	12.73
Type 3	1.06	6.77	8.08	73.57	10.52
Type 4	0.45	13.69	4.84	4.16	76.86

Table 4. The results of the tone recognition rate based on PA-SVM by five types

	Tone recognition rate (%)				
	Type 0	**Type 1**	**Type 2**	**Type 3**	**Type 4**
Type 0	80.76	4.37	5.25	3.50	6.12
Type 1	0.85	84.97	3.41	3.94	6.83
Type 2	0.63	5.94	72.21	10.42	10.80
Type 3	0.84	5.39	6.43	78.96	8.38
Type 4	0.34	10.22	3.61	3.10	82.73

Table 5. The results of the tone recognition rate based on DBN-SVM by five types

	Tone recognition rate (%)				
	Type 0	**Type 1**	**Type 2**	**Type 3**	**Type 4**
Type 0	82.64	3.95	4.73	3.16	5.52
Type 1	0.73	87.14	2.92	3.37	5.84
Type 2	0.54	5.05	76.35	8.87	9.19
Type 3	0.73	4.63	5.53	81.91	7.20
Type 4	0.27	8.29	2.94	2.52	85.98

Though PA-SVM and DBN-SVM use the same features, performance of DBN-SVM is obviously better than that of PA-SVM, and the tone recognition rates of all the five types are improved significantly when the DBN is used to extract high-level features. This might be because that DBN can fuse the multi-source features more efficiently than the shallow models. As shown in Table 6, compared with the P-SVM model and PA-SVM, the proposed DBN-SVM model improves the average rate by 8.61% and 3.07% separately.

Table 6. Comparison of results of the three classifiers

	P-SVM	**PA-SVM**	**DBN-SVM**
Rate (%)	74.42	79.96	83.03

5. Conclusion

This paper presented a DBN-SVM based tone recognition method. In the method, the DBN is employed to integrate prosodic features and articulatory features into high-level tone features. Then, SVM is adopted as the tone classifier. Experiment results on the 863 test set show that the proposed method can increase the accuracy of the tone classification significantly.

Through this study, the potential of deep learning for tone classification can be found. However, the choice of network structure still lacks of perfect theoretical basis. The optimal parameters are often selected by experience or experimental methods, which affects the efficiency of classification. The future research will be how to determine the network parameters of deep learning more quickly and accurately.

Acknowledgement

This paper is funded by the China National Nature Science Foundation (No. 61502150), Key Scientific Research Projects of Universities in Henan (No.19A520004), and Foundation for University Key Teacher by Henan Province (No. 2015GGJS-068).

References

[1] L. W. Cheng and L. S. Lee, "Improved large vocabulary Mandarin speech recognition by selectively using tone information with a two-stage prosodic model," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, 2008, pp. 1137-1140.

[2] H. Wei, X. Wang, H. Wu, D. Luo, and X. Wu, "Exploiting prosodic and lexical features for tone modeling in a conditional random field framework," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, 2008, pp. 4549-4552.

[3] Y. Qian and F. K. Soong, "A multi-space distribution (MSD) and two-stream tone modeling approach to Mandarin speech recognition," *Speech Communication*, vol. 51, no. 12, pp. 1169-1179, 2009.

[4] H. Chao, Z. Yang, and W. Liu, "Improved tone modeling by exploiting articulatory features for Mandarin speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4741-4744.

[5] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, 2006.

[6] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Transactions on Signal Processing Magazine,* vol. 29, no. 6, pp. 82-97, 2012.

[7] S. K. Kim, Y. J. Park, and S. Lee, "Voice activity detection based on deep belief networks using likelihood ratio," *Journal of Central South University*, vol. 23, no. 1, pp. 145-149, 2016.

[8] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3-14, 2017.

[9] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, 2015. pp. 1537-1540.

[10] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771-1800, 2002.
[11] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines (v.3.23)," 2018; http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Hao Chao https://orcid.org/0000-0001-6700-9446

He received his Ph.D. degree in pattern recognition and intelligent system from Institute of Auotmaition, Chinese Academy of Sciences in June 2012. He is currently a lecturer in Henan Polytechnic University. His current research interests include speech signal processing and data mining.

Cheng Song https://orcid.org/0000-0002-6055-6161

He received the M.S. degree in College of Computer Science and Technology, Henan Polytechnic University in 2002. He obtained his Ph.D. from Beijing University of Posts and Telecommunications in 2011. He is currently a lecturer in Henan Polytechnic University. His current research interests include data mining and intelligent information processing.

Bao-Yun Lu https://orcid.org/0000-0001-7717-2861

She received her Ph.D. degree in pattern recognition and intelligent system from Institute of Automation, Chinese Academy of Sciences in June 2011. She is a lecturer in Henan Polytechnic University. Her current research interests include speech signal processing and data mining.

Yong-Li Liu https://orcid.org/0000-0002-0540-865X

He received his Ph.D. degree in computer science and engineering from Beihang University in 2010. He is currently an associate professor in Henan Polytechnic University. His current research interests include data mining and information retrieval.