A Novel Modeling Method for Acoustic Model in Deep Neural Network by Introducing Language Vector

Wei YANG¹, Fu-rong YAN²*, Ke ZHU¹, Cai-jun ZHANG¹ and Xiao-guang HUANG²

¹State Grid Customer Service Center, Tianjin, China
²State Grid Information and Telecommunication Group Co., Ltd. Beijing, China

Keywords: Deep neural network, Acoustic model, Language vector.

Abstract. Much attention has been paid on Multi-language or cross-language speech recognition domain. A kind of speech recognition system with one language and different accents is also an application in the above domain. In this paper, a new concept called language vector is proposed and then is applied in the learning of acoustic model in deep neural network. The proposed method introduces language vector by conditional learning and multi-task learning and dramatically improves the performance of the English speech recognition system aimed at British accent and Chinglish accent.

Introduction

As the increase of computing power and the accumulation of big data, deep neural network (DNN) has increasingly showed its potential power [1] in recent years, especially in the speech recognition field [2,3]. The capability of automatically extracting features for DNN [4] can make its ability of representing the data using features be more suitable for classification tasks. What is more, the output of each hidden layer in a DNN is a nonlinear change of its previous input, the capability makes DNN also learn the hierarchical progressive relationship of features. The above characteristics of DNN make it be noticeably different from the traditional manual marking method [5]. And the stronger representative ability for nonlinear speech makes DNN be superior to the traditional Gaussian Mixture Model (GMM) in speech recognition tasks [6].

With the progress of science and technology, the traditional single-language speech recognition system cannot satisfy the increasingly diversified world, so the multi-language or cross-language speech recognition system is paid more and more attention. However, in contrast with single-language speech recognition system, there are the following questions in the kind of system. Firstly, the acoustic model is less accurate and effective. Secondly, training model takes more time and resource. Thirdly, the robustness to a new scene with mixed speech is much poorer [7,8]. At last, aiming at a special language, whether it is the speaker’s native language directly influences the performance of recognition. Therefore, the speech recognition system with a mixture of native and non-native pronunciation can be regarded as one of the problems which need to be solved by multi-language or cross-language speech recognition system.

Focusing on pure British accent and Chinglish accent, this paper improves the accuracy rate of mixed speech recognition system by strengthening the ability of feature representation of DNN. The ability will be realized by introducing a language vector by conditional learning and multi-task learning.

The paper is organized as follows. In the next section, the concept of language vector is illustrated. Section 3 describes two learning methods including conditional learning and multi-task learning. Section 4 is devoted to establish acoustic model by integrating language vector and DNN based on the above learning methods. Section 5 describes the experimental setup and then analyzes the experimental results. The last section concludes the paper and presents some future directions.
Language Vector

Multi-language or Multi-accents speech recognition system is a kind of application which trains the corresponding model using the mixed speech data of multi-language or multi-accents of a same language and then identifies the corresponding language [9]. The proposed multi-language recognition system can apply in two different accents in English. Therefore, a higher recognition rate of a trained multi-language recognition system means that the system has a stronger ability to distinguish the two types of accents. Based on the above description, the paper proposed a concept of language vector in order to independently extract the capability to distinguish different language or accents and then to improve the performance of recognition. The language vector indicates which language or accent of a same language a kind of speech belongs to. The vector can be multi-dimensional so that the speech locating in a same local vector space comes from a same language or accent.

The whole process of obtaining the model generating language vector is divided into two steps. Firstly, the training speech needs to be divided into frames and each frame will be labeled the pronunciation category by means of the traditional GMM/Hidden Markov Model (HMM). After that, the labeled pronunciation categories are written as three-dimensional language vector (including three situations: native pronunciation, non-native pronunciation and silence) to replace the corresponding state of HMM. All generated language vectors are used to train a small neural network. Similar to the speech recognition system in which DNN is used to replace GMM to be as acoustic model, this paper uses a neural network to generalize the generation ability of language vectors. That is, GMM is a generative model and focuses on reflect the distribution of data and the similarity of the same data, however, DNN is a discriminant model and reflects the differences between the heterogeneous data.

In the paper, a DNN with two hidden layers is chosen as the model of language vector in which the node number of each layer is respectively 1200 and 20. The activation function adopted is respectively linear and sigmoid function, the corresponding expressions are \( f(x) = x \) and \( f(x) = \frac{1}{1 + e^{-x}} \). 40-dimensional FBank features including static, first order temporal derivatives(delta) and second order temporal derivatives (delta-delta) are applied in the paper. Considering the context, 400-dimensional features extracted from 5 neighbor frames before and after the current frame are added. The output node number is 3, corresponding to the dimension of the language vector.

There are the following conclusions through experiments. If 20-dimensional language vector is extracted from the acquired language vector model, that is, the vector is obtained from the output of the second hidden layer before softmax and then is introduced into the training process of the acoustic model. Another, the output of language vector model ultimately is as language vector and the dimension is 3. For the two kinds of language vector, the effect of the former is better than the latter. Additionally, the corresponding language vector of each frame can be independently and several frames belonging a same sentence can share a same language vector. It is found that the performance improvement brought by the latter type is more than by the former through experiments. The average frame of a same sentence inputs into the language vector model, the output corresponds to the common language vector. And, the average frame of a same sentence is obtained by,

\[
F_{\text{average}} = \frac{1}{n} \sum_{i=1}^{n} F_i
\]

where, \( F_{\text{average}} \) and \( F_i \) respectively represent the average feature of frames and the feature of the ith frame in a same sentence.
Conditional Learning and Multi-task Learning

Conditional Learning

For each sample in the training set, some more explicit additional relevant information is added into the input, which will make the feature extraction of DNN be more purposeful and directional. The feature extraction method causes learning of DNN be more effective.

The learning method can be thought of being conducted under conditions, that is called conditional learning [10]. The situation is similar to adding special features into the input, and the key is the way that these features are obtained. If the features are also learned by DNN rather than manually being selected, this means that the output of one DNN is the input of another DNN, which makes the extended feature vector be more adaptable to the later learning of DNN.

What is more, additional relevant information can be added beyond the input layer, such as hidden layers. This can help DNN to more effectively extract features.

Multi-task Learning

Multi-task learning is a transfer learning method, which is to learn the current task and meanwhile to learn other related tasks. So that the performance of single-task learning can be improved by using the information contained in other tasks [11]. The key to multi-task learning is that the tasks must be closely related. The correlation does not mean that the tasks are similar, but rather they share a part of the representation of the data. If the tasks are related but not similar, they can shrink each other's possible function space to improve their own goal generalization ability.

Each hidden layer of DNN is a nonlinear combination of its input, so the upper hidden layer has a more abstract representation than the lower layer, which makes DNN be particularly suitable for multi-task learning [12]. As shown in Figure 1, take two tasks for example, the two tasks share common inputs and several underlying features.

![Figure 1. The diagram of multi-task learning model.](image)

Acoustic Model in Deep Neural Network Introduction Language Vector

Language model and acoustic model are two important parts in speech recognition system. The acoustic model directly determines the performance of speech recognition system. In this paper, the acoustic model is constructed by replacing traditional GMM with DNN. The features of the current frame and their adjacent frames are extracted as the input of the conventional DNN acoustic model. And the output of acoustic model represents the posterior probability of the related frame under the corresponding HMM state.

Different from the usual practice, we use GMM/HMM to obtain the pre-labeled data and then obtain the language vector model. In the end, the language vector is introduced into the training progress of DNN by conditional learning and multi-task learning.
The Introduction of Language Vector by Condition Learning

As an independent feature representation, the language vector can be directly combined with the corresponding data of the training set and then be as the input of DNN. And the language vector can be also separately as the conditional input of several hidden layers. The different modes introducing the language vector by condition learning are shown in Figure 2. Taking 4-layer DNN for example, the language vector is added to the input layer and the first-layer hidden layer respectively.

![Figure 2. The diagram of the language vector introduced into DNN by conditional learning.](image)

The Introduction of Language Vector by Multi-task Learning

Multi-task learning puts multiple related tasks together and learns at the same time. The added tasks will change the dynamic characteristics of the updated weight of network and make the output of classifier mutually influence so as to improve the generalization performance of the classifier.

Constructing an acoustic model and classifying the speech belonging to different language or accents are different, which is the basic condition of multi-task learning when two kinds of tasks study together. As both of them need re-label the data with GMM/HMM, they have a certain correlation. The situation makes it possible that the former’s learning can promote the latter’s learning.

The input of speech recognition and language classification task is speech feature. The paper chooses FBank as the input feature. Therefore, the lower dimensional feature space can be shared by the two tasks. In the paper, a small-scale model is firstly trained using language vector. The trained model contains knowledge which is beneficial for the later speech recognition model and is treated as an initial model, subsequently, it is reused by the later speech recognition model [13]. As shown in Figure 3, the leftmost model is the language vector model which transfers knowledge to a slightly larger scale DNN and uses this DNN as the first several layers of the acoustic model.

![Figure 3. The diagram of the first several layers of the acoustic model adding language vector.](image)

In the experiment, we find that the different number of pre-training hidden layers corresponds to a different degree of transferred knowledge of small-scale network to the hidden layer. Accordingly, the
final improvement performance is also different. Among them, the first three hidden layers are selected as the optimal one, which is taken as an example in this paper.

The Experiment and Analysis

The Experimental Setup

The evaluation experiments are based on the open source tool Kaldi and the provided GPU strategy is used to train the acoustic model. The corpus used in experiments is a mixed speech pool including pure British and Chinglish accent. And in the corpus, 30000 utterances are used for training and 1503 utterances are used for cross validation. The testing utterances contain 503 pure British accent utterances and 1000 Chinglish accent utterances. All pure British accent utterances are collected by European and American students whose native language is English and are recorded by mobile phones. Chinglish accent utterances are collected by Chinese university students without special language training and are recorded by the same microphone. All recordings are wav format with a sampling rate of 16kHz.

At the initial stage of the experiment, the 13-dimensional Mel Frequency Cepstral Coefficients (MFCCs), their delta and delta-delta are used to train GMMs, and then a basic DNN and language vector generation model are trained according to the labeled data obtained by the trained GMMs. At the end, the improved DNN is trained combined with the training of language vector. Additionally, the feature vector used in the training of DNN is a 40-dimensional FBank. Considering the context, 11 frames of symmetrical window are used and the dimension of feature vector is reduced to 200 by linear discriminate analysis (LDA). Both basic DNN and an improved DNN contain 4 hidden layers, each of which has 1200 nodes. For the two networks, the node number of the output is 7552 and the number corresponds to the number of each gaussian model in GMM. And cross entropy is used as the evaluation criterion of the network training. Stochastic gradient descent method is used to train the network and the learning rate is 0.008.

Experimental Results and Analysis

The experimental results are shown in Table 1, word error rate (WER) is used to evaluate the acoustic model.

In Table 1, DNN is an acoustic model trained by the common method and is treated as baseline system. Language vectors are added into the input in DNN-1a and language vectors are introduced to the first hidden layer in DNN-1b. For DNN-2, the first three hidden layers are obtained by pre-training the language vector model, additionally, the frames of a same sentence share a same language vector. It should be noted that, the first affine layer of DNN-1b is taken from the first affine layer of the reference DNN and the weight of the first affine layer is not updated when the language vector is introduced to the first layer hidden layer in order to simplify the experiment. As shown in Table 1, it is found that the introduction of language vector to DNN in both ways improves the recognition performance.

| Model  | WER(%) | Pure British accent | Chinglish accent |
|--------|--------|---------------------|-----------------|
| DNN    | 11.39  | 37.73               |                 |
| DNN-1a | 11.18  | 36.95               |                 |
| DNN-1b | 11.18  | 37.31               |                 |
| DNN-2  | 11.20  | 36.93               |                 |

Conclusion

As global communication becomes more and more extensive, language which is the most convenient mean of communication becomes more and more important, and multi-language or cross-language
recognition system in the field of computer application becomes more and more important. Taking a speech recognition system with multiple accents in a single language as an example, this paper introduces the language vector to the acoustic model of DNN by conditional learning and multi-task learning and the proposed method improves the performance of speech recognition system. In the later stage, we will expand the application range of language vector, such as introducing multiple hidden layers to DNN for language vector and adjusting the weight of the whole neural network at the same time.

Acknowledgement

This research was financially supported by Big Data Application Project Supporting Cloud Sales Business and Research and Application of Key Technology of Intelligent Knowledge Base.

References

[1] Deng, Li, Dong Yu, Deep learning: methods and applications, Foundations and Trends in Signal Processing, 2014, pp. 197-387.

[2] G. Dahl, D. Yu, Li. Deng, A. Acero, Large vocabulary continuous speech recognition with context-dependent DBN-HMMS, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 4688-4691.

[3] A. Mohamed, G.E. Dahl, G. Hinton, Acoustic modeling using deep belief networks, IEEE Transactions on Audio, Speech, and Language Processing, 20 (2012), 14-22.

[4] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science, 313 (2006), 504-507.

[5] Y. Bengio, Learning deep architectures for AI, Foundations and trends® in Machine Learning, 2 (2009), 1-127.

[6] G. Hinton, L. Deng, D. Yu, et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, IEEE Signal Processing Magazine, 29 (2012), 82-97.

[7] J.T. Huang, J. Li, D. Yu, L. Deng, Y.F. Gong, Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 1-5.

[8] D. Yu, L. Deng, Automatic Speech Recognition-A Deep Learning Approach, Springer, October 2014.

[9] H.K. Wang, J. Pan, C. Liu, Research development and forecast of automatic speech recognition technologies, Telecommunications Science, 2 (2018), 1-11.

[10] Z.Y. Tang, L.T. Li, D. Wang D., R.C. Vipperla, Collaborative Joint Training with Multi-task Recurrent Model for Speech and Speaker Recognition, IEEE/ACM Transactions on Audio Speech & Language Processing, 2017.

[11] M.L. Zhang, Research on Multi-Task Learning, Science paper Online, 2011, pp. 1-5.

[12] M.Z. Chen, Research on Chinese and English hybrid recognition based on deep neural network, University of Chinese Academy of Sciences, 2016.

[13] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv:1503.02531v1 [stat.ML], 2015, pp. 1-9.