International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Nexus DNN for Speech and Speaker Recognition

Chittampalli Sai Prakash, J Sirisha Devi

Abstract: ***Over the years, many efforts have been made on improving recognition accuracies on Automatic speech recognition (ASR) and speaker recognition (SRE), and many different technologies have been developed. Given the close relationship between these two tasks, researchers have proposed different ways to introduce techniques developed for these tasks to each other. In this paper an open source experimental framework is proposed for speech and speaker recognition. Then a unified model, Nexus-DNN is developed that is trained jointly for speech and speaker recognition. Experimental results show that the combined model can effectively perform ASR and SRE tasks.***

Keywords: Automatic speech recognition, speaker recognition, Nexus-DNN, Word Error Rate, shared hidden layers

I. INTRODUCTION

Automatic Speech Recognition (ASR) and Speaker Recognition (SRE) are of exceptional hobby in the speech network, and scholars have conducted sizable research on them. The attention of this paper is to explore the relationship between those obligations and to be seeking approaches to improve one's performance with the assist of the opposite. Generally talking, ASR and SRE percentage many similarities in that they're managing equal sort of information and that they use similar statistical / deep mastering fashions. Therefore, a lot studies has been achieved on introducing techniques evolved for them to every other. These include speaker edition, speaker adaptive education and commonplace historical past version for SRE [1] [2].

Most preceding works on this subject matter specializes in enhancing evaluation overall performance on one of the two obligations. Speaker version and speaker adaptive training are strategies for growing speaker-established acoustic models so that higher ASR performance will be performed. Some different research exploits ways to improve SRE overall performance the usage of acoustic models educated for ASR tasks [3]. Just as the manner humans learn how to recognize speech, building a speech recognizer require significant amount of speech data referred to as education data. By exploring styles and traits of those training records, we construct models which could examine from education information to perform transcriptions of unseen information (referred to as test data). The former part of the system is called training, and the latter component is referred to as checking out (or deciphering).

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Mr. Chittampalli Sai Prakash, Department of Computer Science and Engineering Institute of Aeronautical Engineering, JNTU (H) Hyderabad, India
Dr. J Sirisha Devi, Department of Computer Science and Engineering Institute of Aeronautical Engineering, JNTU (H) Hyderabad, India

In a machine studying technique for speech reputation, fashions are basically parameters to be expected from the education statistics. Estimation of these parameters can be performed by means of placing an objective feature over the training information and optimizing the goal. For trying out, we carry out popularity of check facts the use of the educated version, and examine the outputs with ground-reality transcriptions to get an assessment metric.

This paper proposes a joint-version for speech and speaker, Nexus-DNN this is skilled the usage of multi-undertaking studying. Experiments on speech and speaker reputation are provided and the outcomes are in comparison with baseline structures. It is shown that this Nexus-DNN model is effective in utilizing restrained quantity of education facts for ASR and speaker popularity.

In section II, the proposed model for speech and speaker recognition is presented. In section III, a step-by step process of recognition using Nexus- DNN is briefly explained. Followed by section IV and V in which experimentation and results are discussed.

II. PROPOSED MODEL OF SPEECH AND SPEAKER RECOGNITION

Acoustic models are trained for ASR can successfully be used to enhance speaker reputation overall performance, and speaker i-vectors primarily based speaker model is effective in improving ASR accuracy. In this phase, speech and speaker reputation are combined to discover a answer that tackles both duties on the equal time. The connections among speech and speaker recognition has long been identified by means of researchers. However, compared with the efforts spent on speaker model and speaker popularity, fewer attempts had been made on joint modeling of speech and speaker [4]. There are many reasons leading to this:

1. Most studies projects are set up to observe one specific trouble in place of tackling two or greater of all of them together.
2. Focusing on ASR or speaker recognition in my view usually yields higher effects in comparison to multi-tasking.
3. Speech facts units are commonly designed to deal with either ASR or speaker popularity, which makes it difficult to behavior research on joint modeling.

Despite of most of these reasons, joint-modeling it is an crucial topic to take a look at. Firstly, the human mind is capable of carry out two or greater undertaking simultaneously the use of subconscious mind. When picking up telephone calls, or paying attention to TV shows / radios, people generally tend to apprehend the speaker and contents at the identical time [5].

Secondly, ASR and speaker reputation can be beneficial to each other, so it's far natural to suppose that a joint version may want to advantage from mastering to carry out both duties. In a current paper posted by using Google Research, the authors proposed a deep learning model to perform 8 obligations on the identical time. This motivates us to searching for higher nexus modeling methods for speech [6].

In this paper a novel method based on Nexus-DNN is proposed which differs from the traditional models in the following aspects:

- It focuses on conversational ASR and text-independent speaker recognition.
- It is a joint neural network model rather than a combination of different models.
- It is evaluated on standard test sets for both ASR and speaker recognition.

III. JOINT MODELING OF SPEAKER AND SPEECH

A. General Design of Nexus-DNN

The Nexus-DNN version for ASR and speaker reputation takes segments of speech features as inputs, passes them through some of shared hidden layers, and then separates out into sub-networks that predict HMM states and speaker identification respectively. The structure of the proposed model is proven in Figure 1. A pooling layer is placed in the SRE subnet to average out sequence of activations. These pooled activations are then exceeded directly to predict speaker identity. During checking out, the ASR subnet can be used to generate body log-likelihoods for WFST decoder, and the SRE subnet may be used to generate speaker embeddings for speaker popularity, referred to as jd-vector. Since the pooling layer reduces the dimension of activations, there might be a size mismatch between layers before and after pooling. To update the network in a mini batch fashion, unique cares should be taken whilst making ready the records. To be unique, input information for the network are packed into a three-dimensional matrix of size [num batches, segment length, feature dim], and activations of all DNN layers earlier than the pooling layer are of length [num batches, segment length, hidden units]. The pooling layer take averages of hidden activations over speech segments, so the output of the pooling layer becomes [num batches, hidden units]. The labels for ASR subnet and SRE subnet are of sizes [num batches, segment length, num states] and [num batches, num speakers] respectively.

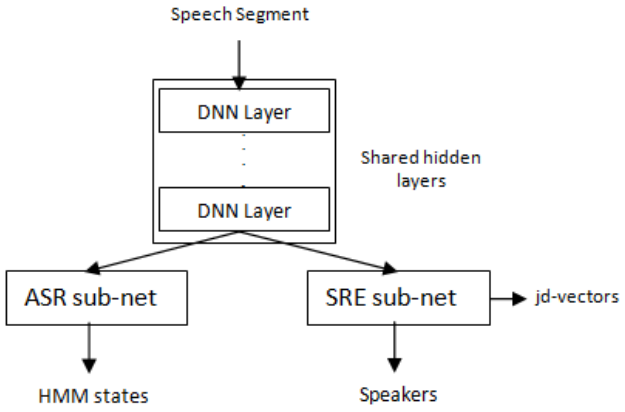

Figure 1. Structure of Nexus-DNN model

B. Data preparation

Since joint modeling requires data labeled with both text and speaker information, only limited choices are available for model training. Here in this project, the TIMIT data set is chosen as the main corpus, which contains 19k utterances from various speakers. These utterances are randomized and saved sequentially on disk before training starts. Filter-bank features are used as input features in this project as they are low level representation of speech signals compared to other well-developed ASR or speaker recognition features.

During training, blocks of utterances are loaded into memory one at a time. The data generator packs utterances into batches before sending them over for model back propagation. For utterances longer than a pre-defined segment length L, they are broken into pieces using a sliding window. For those that are shorter than L, zeros are padded to the end of those utterances. The data generator also needs to prepare ASR labels and speaker labels along the way.

Special care must be taken regarding silence inside speech utterances. Since silence frames are not useful for speaker recognition, they must be excluded during pooling. To achieve this operation, a mask is prepared for each segment using pre- computed voice activity detection information.

C. Loss function

The loss function for model training is defined by:

$$L(\theta) = -\sum_{s=1}^{S}\sum_{t=1}^{sT} h_{s,t} \log P(h_{s,t}|O_{s,t}) - \beta \sum_{s=1}^{S} x_s log P(x_s|O_s)$$

which is an interpolation of cross-entropy losses for ASR and speaker recognition. Here $h_{s,t}$ denotes the HMM state for frame t of speech segment s, and $o_{s,t}$ is the observed feature vector that corresponds to $h_{s,t}$, x_s is the correct speaker for segment s and o_s is speech features for segment s. β is the interpolation weight.

D. Making predictions

To evaluate this model, ASR and speaker recognition are tested on standard data set. For ASR decoding, the ASR branch of the network are used to generate frame log-likelihoods. These log-likelihoods are passed into Kaldi's WFST decoder via a pipe to generate decoding outputs. For speaker recognition, activations after the pooling layer are collected as speaker embeddings, just as the way x-vector is generated. These speaker embeddings, referred to as jd-vector, is used for scoring methods, like cosine scoring or PLDA scoring.

E. Buckets for training and testing

To ensure SRE subnet generalizes well to speech segments of different lengths, bucket training is implemented for x-vector. During data preparation, speech segments for training are fed into buckets of different sizes. Then, in training phase, speech segments in the same bucket are passed into the model trainer in batches to perform an SGD based model update. During model evaluation, buckets are also used to generate jd-vectors for speech segments of different sizes.

IV. EXPERIMENTAL SETUP

The Nexus-DNN model used has 3 shared hidden layers with 2048 hidden nodes per layer. For the ASR subnet, 3 more hidden layers with 2048 nodes per layer are used, before a final softmax layer of 5238. For the SRE subnet, the number of hidden nodes is projected down to 1500 before pooling. The layer after the pooling layer further reduces number of hidden nodes to 512, before sending activations to a final softmax layer. Jd-vectors are extracted after the pooling layer, which has 512 hidden nodes.

The TIMIT data set [7] is used for model training. To make sure the pooling operation in SRE subnet is stable during training, utterances that are 2 seconds or longer are preselected, which leaves 158k utterances from 520 speakers, totaling 270 hours of speech. Histograms of utterance lengths and utterances per speaker are shown in Figure 2.

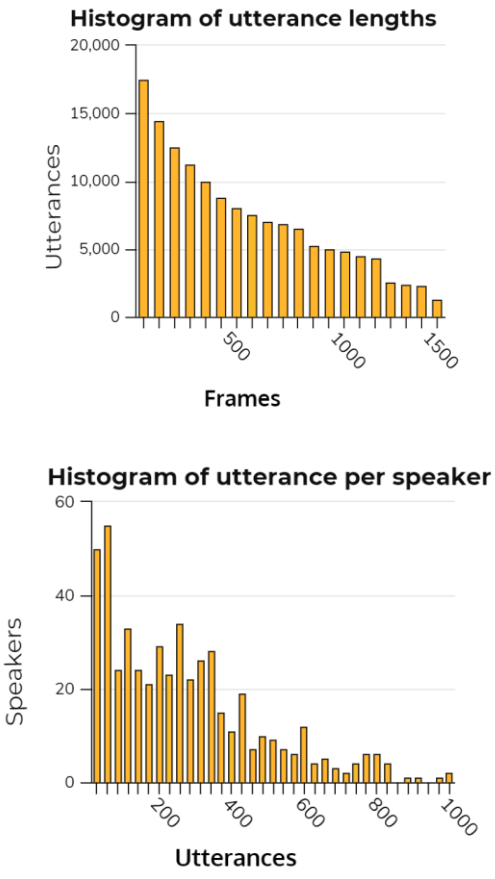

Fig. 2. Histograms of utterance lengths and utterances per speaker

Standard filter-bank features (24-dimension) are extracted, and then expanded to 264 dimensions to include context information by concatenating features of neigh- boring frames. Sliding window mean normalization is performed on the features before a global mean and variance normalization. Alignments for ASR subnet are generated using a baseline GMM-HMM system. This baseline system is built following the standard Kaldi for TIMIT data set, where LDA, MLLT and SAT techniques are used during model training. New-bob method is used for learning rate scheduling, and standard SGD is used for model update.

For model evaluation, the NIST 2000 Hub5 evaluation set (Eval2000) [8] is used for ASR experiments, and SRE 2010 data set [9] is used for speaker recognition experiments.

V. RESULTS AND ANALYSIS

A. SRE performance

Testing of speaker recognition is evaluated on the SRE 2010 data set. Table 1 compares recognition performances of different systems on the test set. As is shown in the table, the baseline i-vector system performs best in terms of EER. This is because the training data is very limited in this experiment - only around 300 hours in total. The i-vector model, as a generative one, usually performs better than discriminative models when limited data are available. This observation is also compatible with that published in [10]. The Kaldi x-vector system and the TIK x-vector system give comparable results regarding EER. The Nexus-DNN implementation, however, outperformed both x-vector systems and achieves an EER of 5.30 under LDA+PLDA scoring. This result shows that phonetic content could help the Nexus-DNN to model speaker embeddings.

Table 1. EER of Nexus-DNN model for speaker recognition

Baseline DNN	16.1
Nexus-DNN (beta 0.01)	15.36

Table 2. WER of Nexus-DNN model for speech recognition

	cosine	**LDA**	**PLDA**	**LDA+PLDA**
Baseline i-vector	22.31	5.23	5.51	4.93
Kaldi x-vector	38.54	11.45	8.94	8.28
TIK x-vector	37.77	18.65	8.16	8.41
TIK jd-vector (beta0.01)	36.21	7.68	5.78	4.43

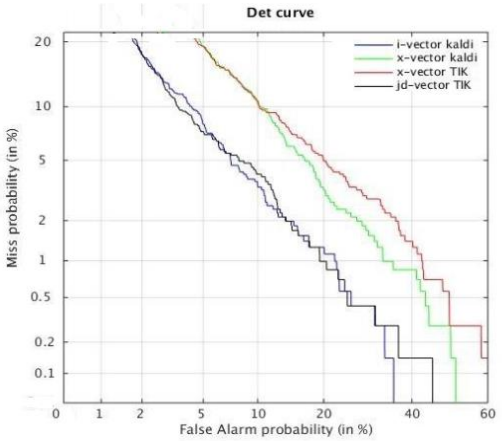

Fig. 3. DET curve for speaker recognition performance

The DET curve of all four systems on SRE 2010 test set as shown in figure 3. As shown in the figure, performances of i-vector and jd-vector systems are quite close to each other. However, in the low miss-rate region, the i-vector system becomes better.

In general, jd-vector performs better than both x-vector systems.

Table 3. ASR and SRE performances of models trained with different βs

	Validation		Evaluation	
beta	**Frame acc (%)**	**Speaker acc (%)**	**EER (%)**	**TIMIT Word Error Rate (WER%)**
0.1	28.03	98.24	3.32	15.91
0.01	28.17	97.01	3.16	15.36
0.001	27.43	81.15	8.19	18.49
0.0001	25.38	54.25	15.04	18.32

B. ASR performance

Testing of speech recognition is evaluated on an Eval2000 dataset [11]. Currently, we do not observe recognition improvement over the baseline DNN model. Adding additional targets of speaker labels introduces slight negative influences on the speech recognition part.

C. Adjusting Beta

As is covered in the last section, the loss function for training Nexus-DNN model is a weighted sum of ASR loss and SRE loss. Here β is used to adjust the weight placed on speaker recognition.

In the Table 3 EER of Nexus-DNN model with different interpolation weight Validation columns show frame accuracy and speaker accuracy on cross-validation set when the last iteration of model training is finished. Evaluation columns show EER on SRE 2010 test set and WER on Hub5 Eval 2000 TIMIT portion.

As shown in the table, the SRE portion general performs better as β becomes bigger. The best SRE performance is achieved when β is set to 0.1. However, as β gets bigger, the ASR performance begins heading worse. The trade-off between ASR and SRE performances can be balanced by adjusting β.

VI. CONCLUSION

Experiments show that Nexus-DNN model is effective in using a limited amount of training data for a neural network based speaker recognition model. On ASR, the Nexus-DNN performs comparably to baseline DNN systems without feature transformations. The Nexus-DNN model, being a feed forward network, is much faster during training, but a bit weaker in utilizing high level representations. The trade-off between model complexity and training speed is worth more investigation.

REFERENCES

1. Qian, Y., Yin, M., You, Y., and Yu, K. (2015). Multi-task joint-learning of deep neural networks for robust speech recognition. In Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on, pages 310{316. IEEE.
2. Saon, G., Soltau, H., Nahamoo, D., and Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In ASRU, pages 55-59.
3. Seltzer, M. L. and Droppo, J. (2013). Multi-task learning in deep neural networks for improved phoneme recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 6965-6969. IEEE.
4. Su, H. and Xu, H. (2015). Multi-softmax deep neural network for semisupervised training. In Sixteenth Annual Conference of the International Speech Communication Association.
5. Tang, Z., Li, L., and Wang, D. (2016). Multi-task recurrent model for speech and speaker recognition. In Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Paci_c, pages 1-4. IEEE.
6. Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pages 4960{4964. IEEE.
7. Sunnydayal V., Sirisha Devi J., Nandyala S.P. (2019) Hybrid Method for Speech Enhancement Using α-Divergence. In: Saini H., Sayal R., Govardhan A., Buyya R. (eds) Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, vol 74. Springer, Singapore
8. Fiscus, J., Fisher, W. M., Martin, A. F., Przybocki, M. A., and Pallett, D. S. (2000). 2000 nist evaluation of conversational speech recognition over the telephone: English and mandarin performance results. In Proc. Speech Transcription Workshop. Citeseer.
9. J. Sirisha Devi ; Srinivas Yarramalle ; Siva Prasad Nandyala ; P. Vijaya Bhaskar Reddy (2017) 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)
10. Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In Proc. Interspeech.
11. Li, L., Tang, Z., Wang, D., Abel, A., Feng, Y., and Zhang, S. (2017). Collaborative learning for language and speaker recognition. In National Conference on Man-Machine Speech Communication, Springer.

AUTHORS FROFIEL

Chittampalli Sai Prakash, Is Pursuing Under Graduation In Computer Science Discipline At Institute Of Aeronautical Engineering. He Is Good At Programming Skills Like Python, Javascript, C. He Is Proficient In Maintaining The Computer System And Existing Software.

Dr. J Sirisha Devi, Was Awarded B. Tech. In Computer Science And Engineering From Acharya Nagarjuna University -2003. She Was Awarded M. Tech. In Computer Science And Engineering From GITAM University, Visakhapatnam - 2010. She Was Awarded Doctorate In The Year 2016. Her Research Interests Include Human Computer Interaction And Natural Language Processing