Incorporating Speaker Normalizing Capabilities to an End-to-End Speech Recognition System

Hari Krishna Vydana, Sivanand Achanta, Anil Kumar Vuppala

Speech Processing Lab, KCIS
International Institute of Information Technology, Hyderabad, India
{hari.vydana,sivanand.a}@research.iiit.ac.in, anil.vuppala@iiit.ac.in

Abstract

Speaker normalization is one of the crucial aspects of an Automatic speech recognition system (ASR). Speaker normalization is employed to reduce the performance drop in ASR due to speaker variabilities. Traditional speaker normalization methods are mostly linear transforms over the input data estimated per speaker, such transforms would be efficient with sufficient data. In practical scenarios, only a single utterance from the test speaker is accessible. The present study explores speaker normalization methods for end-to-end speech recognition systems that could efficiently be performed even when single utterance from the unseen speaker is available. In this work, it is hypothesized that by suitably providing information about the speaker's identity while training an end-to-end neural network, the capability to normalize the speaker variability could be incorporated into an ASR system. The efficiency of these normalization methods depends on the representation used for unseen speakers. In this work, the identity of the training speaker is represented in two different ways viz. i) by using a one-hot speaker code, ii) a weighted combination of all the training speakers identities. The unseen speakers from the test set are represented using a weighted combination of training speakers representations. Both the approaches have reduced the word error rate (WER) by 0.6, 1.3% WSJ corpus.

Index Terms: Automatic speech recognition system, end-to-end speech recognition system, speaker normalization, transform based normalizations, one-hot speaker codes

1. Introduction

Hidden-Markov model-Gaussian mixture model (HMM-GMM) based acoustic models are the most widely used acoustic models in traditional automatic speech recognition systems (ASR). HMM-GMM based ASR is a modular system comprising of modules such as feature processing, speaker normalization, decision trees, acoustic model etc and the extensive study of these modules over the past decades have resulted in improved performances [1]. The hybrid models have a downside that, the objective function to which the hybrid models are optimized is quite different from the true performance measure (Sequence level transcription accuracy) [2, 3, 4, 5]. The development of connectionist temporal classification (CTC) objective function and attentional mechanism have put-forth end-to-end training reducing the mismatch between the objective function that is optimized during the training and true performance measure [3, 6]. The capability to train deeper networks have enriched the capability of training deep neural networks to learn complex nonlinear functions [7]. These end-to-end systems have transformed ASR from a modular approach to a holistic approach, and reducing the mismatch between the end goal of the system and the objective function to which it is optimized during the training [3, 6, 8]. In this perspective, there is a need to study and re-investigate various modules such as feature processing, speaker normalization etc, for attaining the performance gains to an end-to-end speech recognition systems.

Speaker normalization is a crucial module in operating an ASR with high performance. Though there have been several approaches for speaker normalization in hybrid ASR, an optimal mechanism of speaker normalization in an end-to-end speech recognition systems is largely unexplored. In this study, we explore approaches for speaker normalization in the perspective of end-to-end speech recognition system. The key role of speaker normalization is to reduce the performance drop when speech recognition system encounters an unseen speaker [9]. The degradation in the performance of speech recognition systems due to speaker variabilities is addressed by two major approaches, the former approach is to reduce the speaker variabilities by speaker normalization [10, 11, 12] and later is by speaker adaptation [13, 14]. The variabilities that are specific to a speaker are normalized to develop a speaker independent recognition system and the developed system can be adapted to operate for a specific speaker. The remaining paper is organized as follows, section 2 describes some of the related works. Section 3 describes the database and the experimental setup used in this work. The details of the proposed approach for speaker normalization is described in section 4. Complementarity of the proposed approach with fMLLR based normalization approach is explored in section 5. Summary and future scope are presented in section 6.

2. Related work

Some of the widely used speaker normalization approaches are Vocal tract length normalization scheme (VTLN) [11, 12] and Feature space maximum likelihood regression (fMLLR) [10]. VTLN approach aims to warp the frequency axis to account for the varying vocal tract lengths across the speakers and the warp factor is computed from the speaker's data. VTLN warp factor for a speaker can also be computed through grid search to choose the optimal warping factor to maximize the likelihood of speakers data for the corresponding label sequence [12]. Speaker normalization using fMLLR is carried out by estimating a linear transform over the input features that would maximize the likelihood of speakers data for the given label sequence, during the test time an initial label sequence is obtained for the existing model and the features from the speaker are transformed to maximize the likelihood [10]. Cepstral mean and variance of the data normalized on per speaker basis is used as a kind of speaker normalization procedure.

Majority of the approaches model the speaker characteristics as a linear transform over input features or certain model parameters and use them to normalize the speaker-specific char-

10.21437/SLTU.2018-36

acteristics. The performance of such speaker normalization hugely relies on the amount of data available to compute the transform, less data per speaker results in a poor estimate of transform and poor normalization [10]. In a practical scenario, only a single utterance of an unknown speaker can be accessed and the speaker normalization approach should be able to normalize the speaker variabilities and yield a better performance using the single available utterance. In VTLN and fMLLR normalization methods, every speaker is modeled independent of the remaining speakers and such approaches can not model the similarity and variability across the speakers [15]. In the proposed approach, we hypothesize that by suitably providing information about the identity of the speaker along with features, the capability of speaker normalization can be embedded into a neural network that is trained for speech recognition. Where the performance of this normalization approach relies on efficiently representing the characteristics of unseen speakers from the characteristics of trained speakers.

Some of the related studies for speaker normalization in a hybrid ASR are briefly described, spectral features are appended with i-vectors and the appended features are explored for training hybrid ASR [16]. In [17], the bottleneck features accumulated over the entire speaker is appended with spectral features and the appended features have been used as speaker normalized features. Some additional weights are learned at every layer using i-vectors, and the network is used to develop a speaker adapted hybrid ASR. In [18], two bottleneck networks were trained to classify speaker and phone identities, these bottleneck features are appended and the appended features are used to train a deep neural network for speech recognition. In an end-to-end ASR, the acoustic model used is sequential as opposed to a deep neural network (DNN) in hybrid ASR. The end-to-end networks are expected to benefit more from the speaker information locally and the accumulation of local information can happen in the sequential acoustic model, where utterance level speaker representations have proven to perform better in the case of a hybrid ASR.

This work proposes a new speaker normalization method, which aims at representing the characteristics of test speaker as a weighted combination of the set of known speakers on which ASR has been trained. Such system would demand a single unifying framework to normalize over multiple speakers. In this work, representational power of DNN is used for modeling the speaker characteristics of unseen speakers.

3. Database & Experimental setup for end-to-end speech recognition system

3.1. Database

Wall street journal corpus (WSJ) comprising 80 hrs of speech data has been used during the study [19]. The utterances from si-284 are used for training the speech recognition systems and the performance is evaluated on eval-92. The training corpus comprises of data from 282 speakers, while testing corpus comprises of 8 speakers that are independent of the train set. The standard evaluation data (eval 92) comprises of 333 utterances i.e., approximately 40 utterances per speaker.

3.2. Architecture

In this work, Recurrent neural network (RNN)-CTC based speech recognition systems have been explored. The network has 4 layers of bidirectional long short-term memory networks (BLSTMs) and each layer is comprising of 360 units. In this work, a batch size of 10 is used. During the training, forward propagation is done with a batch of 10 utterances and backpropagation is done for every utterance. Learning rate is reduced by a factor of 0.5 when a decrease in validation accuracy is encountered. The training is progressed till a minimum increase of 0.05 is maintained in validation accuracy. In this work, RNN-CTC is optimized to generate character sequences from the input acoustic sequences. The character-based lexicon generated from CMU dictionary and NIST trigram language model has been used in this work. During the study, spectral features with deltas and delta-deltas are used as a feature representation. The gradient at each layer is clipped at a maximum value of 50, all the parameters of the model are initialized from a uniform distribution between [-0.1, 0.1].

4. Proposed approach for speaker normalization

4.1. Approach 1

Widely used speaker normalization methods can be generalized as the linear transform over the input data computed per speaker. The hypothesis for the proposed approach is that, by suitably giving information about the speaker's identity while training the model, the model would be optimized such that the speaker normalization occurs implicitly. The efficiency of the approach relies on the form in which information about training and unseen speaker's identity is presented to the network. In this work, the identity of speaker in the train set is represented using a one-hot vector. The one-hot vectors of the speaker are appended with spectral features and the appended features are used for training the RNN-CTC model. In this approach, the unseen speaker characteristics are to be represented by a weighted combination of known (trained) speakers. A speaker recognition system trained to classify the training speakers (SPKID-DNN) is employed for this task. For the test utterance, the softmax vectors obtained from the SPKID-DNN are appended with the spectral features and the appended features are used to test the speech recognition system. The softmax vectors from SPKID-DNN is expected to represent the unseen test speaker as a weighted combination of known (training) speakers. The performance of the proposed approach is presented in Table 1.

In this study, SPKID-DNN is a fully connected feed forward deep neural network comprising of 4-hidden layers with rectified linear units (ReLU) and a softmax output layer. The architecture of SPKID-DNN is 39R-1024R-1024R-1024R-1024R-282S. During the study, SPKID-DNN needs to be generalizable over the unseen speakers, to obtain such a network various SPKID-DNNs have been trained with different sized datasets i.e., full, 10, 20, 30, and 60 utterances per speaker from the training set and 10 utterances from each speaker is used as validation set for training SPKID-DNN. Spectral features are used to train SPKID-DNN. The network is optimized using Adadelta [20] optimizer to minimize the cross-entropy loss, an initial learning rate of 0.001 is used. Learning rate is reduced by a factor of 0.5 upon encountering a decrease in validation accuracy. The network is trained until a decrease in validation accuracy is observed over three successive epochs.

Row 1 of Table 1 are various speaker normalization approaches and row 2 are the word error rates (WERs) of ASR obtained using the corresponding normalization. Column 2 of Table 1 is WER of ASR with fMLLR speaker normalization. The reported WER is on the eval-92 set comprising of 8 speak-

Table 1: *Performance of the proposed speaker normalization approach using one-hot speaker codes to represent the identity of the speaker (WER-word error rate).*

Speaker normalization method	fMLLR	fMLLR per utt	MVN per spk	MVN per utt	SPKID -DNN-10utt	SPKID -DNN-20utt	SPKID -DNN-30utt	SPKID -DNN-60utt	SPKID -DNN-full	SPKID -DNN30utt -sent label
WER	8.83	11.34	9.92	10.17	9.66	9.48	9.41	9.45	9.55	9.85

ers and each speaker has around 40 utterances. WER in Column 2 is obtained by using data from all 40 utterances in estimating the fMLLR transform. Column 3 is the WER obtained when data from single utterance is used to estimate fMLLR. Form Column 2 and 3 it can be observed that the performance of the systems degrades with insufficiency in the data for estimating fMLLR transform. Column 4 is the system trained using the features with mean and variance normalization (MVN) performed on per speaker basis. Column 5 is the ASR system trained with mean and variance normalization performed using the data from single utterance (MVN-per utt). In the proposed approach, the identity of the training speaker is presented as on-hot vectors. Spectral features are appended with a one-hot speaker vectors of 282 dimensions to form 321-dimensional input features and these features are used to train an RNN-CTC network. The performance of the proposed approach evaluated using various SPKID-DNNs is presented in column 6-9 in Table 1. From columns 6-9, it can be noted that the SPKID-DNN-30utt have performed better compared to other systems. From Table 1, the ASR system with the proposed normalization (i.e., trained with one-hot speaker vector during the training and representing unseen speaker as a weighted combination of trained speakers) have reduced the absolute word error rate by 0.6. The approach is more practically feasible as it used only single utterance to normalize the speaker, unlike fMLLR which uses around 40 utterances for normalizing speaker variabilities. For the given test utterance, the closest speaker from the train speaker is found using SPKID-DNN and the corresponding one-hot vector is used for testing and the obtained WER is given by column 10 of Table 1. Though speaker is an attribute of the entire utterance, in this work a DNN is preferred as it would account for the local variabilities and accumulation of these representations over the utterance is left to the sequential acoustic model (RNN-CTC).

4.2. Approach 2

Approach-1 implicitly assumes that every speaker is independent of the other speakers in the training set. Such an assumption can not model inter-speaker similarity and variability, which might be helpful in modeling the unseen speaker. Moreover, approach-1 cannot model the phonetic context along with the speaker's identity. To embed this additional information, the softmax vectors of SPKID-DNN is used as a speaker vector. The speaker vector is appended with spectral features and the appended feature representation is used for training RNN-CTC model. Block diagram of speaker normalization using approach 2 is presented in Fig.1. The WERs obtained from the proposed approach are reported in Table 2. In this work, multiple ASRs have been trained using different SPKID-DNNs to generate speaker vectors.

Table 2: *Performance of the proposed speaker normalization approach using softmax vectors from SPKID-DNN as representations for the speakers identity.*

Speaker normalization method	SPKID -DNN -10utt	SPKID -DNN -20utt	SPKID -DNN -30utt	SPKID -DNN -60utt	SPKID -DNN -full	i-vector	SPKID -DNN -Bottleneck
WER	9.84	9.21	**8.88**	9.14	8.93	10.97	10.10

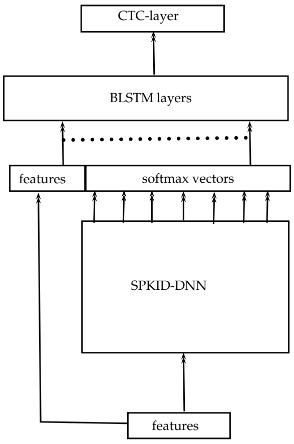

Figure 1: *Block diagram of the proposed speaker normalization approach described as Approach-2 in section. 4.*

From Table 1 and 2 it can be observed that, though the proposed speaker normalization method uses only a single utterance it can yield a performance that is equivalent to using 40 utterances in computing a fMLLR transform. The proposed method is more advantageous and employing the proposed speaker normalization approach an absolute reduction of 1.3 in WER can be observed.

Row 1 of Table 2, is the speaker normalization approach and row 2 is the WER of ASR trained with the corresponding speaker normalization. Column 2-6 are the WERs of ASR systems trained using speaker normalization method proposed in approach-2. Column 8 is the ASR system trained by appending 100-dimensional i-vector along with spectral features and the i-vectors used in this work are computed using a Universal background model (UBM) of 512 components, all the training data is used to compute i-vectors. Pre-softmax activations of SPKID-DNNs is used as speaker vector and appended with the spectral features and the performance of the ASR systems is presented in column 8. The performance of ASR is superior when softmax vectors from SPKID-DNN-30utt are used training the network. Having some utterances for training ASR systems which are unseen by SPKID-DNN have helped the model to generalize over the unseen test set.

During the study, recurrent neural networks (RNNs) have been explored to classify the training speakers (SPKID-RNN). Deep bi-directional LSTMs comprising of three layers and each layer comprising of 250 units have been used to train the SPKID-RNN. The network is optimized as mentioned in section .4.1. The speaker vectors obtained from SPKID-RNN are used in the proposed approach and the performances of ASR are tabulated in column 2,3 of Table 3. The speaker vectors obtained from the SPKID-DNN are appended with speaker vectors SPKID-RNN are combined speaker vectors are used in the proposed approach the performance of the ASR using the combined speaker vector is presented column 4,5 of Table 4. From Table 3, it can be observed that the speaker vectors from SPKID-DNN have performed better that SPKID-RNN. It can also be noted that the performance of proposed approach using speaker vectors from both SPKID-DNN and SPKID-RNN have performed similar to using only SPKID-DNN.

Table 3: *Performance of the proposed speaker normalization approach using softmax vectors from SPKID-DNN and SPKID-RNN as representations for the speakers identity.*

Speaker normalization method	SPKID -LSTM -30utt	SPKID -LSTM -fullutt	SPKID -LSTM-DNN -fullutt	SPKID -LSTM-DNN -30utt
WER	11.48	9.82	11.02	**8.90**

From Table 1 and 2 it can be observed that, though the proposed speaker normalization method uses only a single utterance it can yield a performance that is equivalent to using 40 utterances in computing a fMLLR transform. The proposed method is more advantageous and employing the proposed speaker normalization approach an absolute reduction of 1.3 in WER can be observed.

5. Complementarity of the proposed approaches with fMLLR based speaker normalization

As fMLLR is a global transform to normalize the speaker variabilities and the proposed approach is expected to account for local variabilities. To study the complementary nature of the proposed approach along with fMMLR based normalization. Both the proposed normalization methods are employed while training an ASR with fMLLR features. The performances obtained by combining both the normalizations are reported in Table 4. In this work, the SPKID-DNN is also trained with fMLLR features.

Table 4: *Performance of the speech recognition systems developed using fMLLR features and the proposed normalization approaches.*

Speaker normalization method	one-hot speaker codes		SPKID-DNN softmax vectors	
	fMLLR-SPKID -DNN-30utt	fMLLR-SPKID -DNN-full	fMLLR-SPKID -DNN-30utt	fMLLR-SPKID -DNN-full
WER on eval-set	8.75	8.77	**8.59**	8.93

Form Table 4, the complementary nature of the proposed approach can be observed, a slight decrease in WERs is observed. In this study, SPKID-DNNs trained with spectral features can also be used and they have also yielded similar performances.

6. Summary & conclusion

This work explores speaker normalization methods that would efficiently normalize the speaker variabilities using a single utterance from the test data. In this work, it is hypothesized that by suitably describing the identity of the speaker while training an end-to-end neural network, the model is optimized such that it is robust to speaker variabilities. During the study, the identity of the known speaker is represented by using two approaches viz., one-hot speaker codes and a weighted combination of training speakers, while the un-seen speaker's identity is represented as a weighted combination of the seen speakers representations. For an utterance, the weighted combination of training speakers is obtained by a DNN trained to classify the training speakers. This normalization method would be effective even for single test utterance and the proposed approaches reduce the absolute WER by 0.6 and 1.3 respectively.

7. ACKNOWLEDGEMENTS

The authors would like to thank DeitY for supporting the research under the Visvesvaraya PhD fellowship scheme.

8. References

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks.," in *Proc. International Conference on Machine Learning*, 2014, vol. 14, pp. 1764–1772.

[3] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 2016, pp. 4945–4949.

[4] Liang Lu, Xingxing Zhang, Kyunghyun Cho, and Steve Renals, "A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition.," in *Proc. INTERSPEECH*, 2015, pp. 3249–3253.

[5] Liang Lu, Lingpeng Kong, Chris Dyer, Noah A. Smith, and Steve Renals, "Segmental recurrent neural networks for end-to-end speech recognition," in *Proc. INTERSPEECH*, 2016.

[6] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. International Conference on Machine Learning*. ACM, 2006, pp. 369–376.

[7] Yoshua Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[8] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 2016, pp. 4960–4964.

[9] Lahiru Samarakoon and Khe Chai Sim, "Learning factorized feature transforms for speaker normalization," in *IEEE Workshop on. Automatic Speech Recognition and Understanding*, 2015, pp. 145–152.

[10] Phil C Woodland, "Speaker adaptation for continuous density hmms: A review," in *ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, 2001.

[11] Li Lee and Richard C Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1996, vol. 1, pp. 353–356.

[12] Li Lee and Richard Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, 1998.

[13] Yajie Miao, Hao Zhang, and Florian Metze, "Speaker adaptive training of deep neural network acoustic models using i-vectors," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1938–1949, 2015.

[14] Shaofei Xue, Ossama Abdel-Hamid, Hui Jiang, Lirong Dai, and Qingfeng Liu, "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1713–1725, 2014.

[15] Hagai Aronowitz, Dror Irony, and David Burshtein, "Modeling intra-speaker variability for speaker recognition," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[16] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors.," in *IEEE Workshop on. Automatic Speech Recognition and Understanding*, 2013, pp. 55–59.

[17] Hengguan Huang and Khe Chai Sim, "An investigation of augmenting speaker representations to improve speaker normalisation for dnn-based speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 2015, pp. 4610–4613.

[18] Marc Ferras and Hervé Bourlard, "MLP-based factor analysis for tandem speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 2013, pp. 6719–6723.

[19] John Garofalo, "Wall street journal-based continuous speech recognition corpus," *Linguistic Data Consortium, Philadelphia*, 1994.

[20] Matthew D. Zeiler, "Adadelta: An adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.