Synthetic Speaker Models Using VTLN to Improve the Performance of Children in Mismatched Speaker Conditions for ASR

D. R. Sanand and T. Svendsen

Norwegian University of Science and Technology, Trondheim, Norway

[ramad,torbjorn]@iet.ntnu.edu

Abstract

The paper proposes to train synthetic speaker models using vocal tract length normalization (VTLN). Speaker adaptation based approaches require certain amount of data from the test speaker to either update or transform the model parameters of the trained model. If there is very little or no data available from the test speaker, we propose to create a synthetic speaker model that is acoustically close to the test speaker by scaling the training data with VTLN. For this purpose, we train multiple VTLN warped speaker independent (SI) models by scaling the training data with VTLN and choosing one of the models that is acoustically close to the test speaker for performing recognition. We show that the proposed approach is advantageous in mismatched speaker conditions, especially while recognizing children speakers using models trained on adult speech.

Index Terms: Synthetic speaker models, Vocal tract length normalization, Speaker adaptation, Speech recognition.

1. Introduction

The paper primarily focuses on improving the performance of children speech recognition using models trained on adult speech. Due to limited availability of children data for training, it is quite common to employ models trained on adult speech for recognizing children. Since there are a wide variety of differences in acoustic and linguistic characteristics between children and adults [1], the performance of the system degrades drastically while recognizing children using models trained on adult speech [2, 3, 4]. A wide variety of techniques have been proposed in literature to reduce speaker variability in ASR, that include vocal tract length normalization (VTLN) [5, 6], maximum a-posteriori (MAP) adaptation [7], maximum likelihood linear regression (MLLR) [8] and constrained MLLR (CMLLR) [9, 10]. These techniques are applied as stand alone operations or in combination to improve the performance in either matched or mismatched speaker conditions [11, 12, 13].

The paper proposes to train synthetic speaker models using VTLN. Adaptation based approaches transform the parameters of the SI model to create a speaker dependent model that is acoustically closer to the test speaker. They require certain amount of data from the test speaker to modify the model parameters. If there is very little or no data available from the test speaker, it is difficult to perform adaptation. The paper proposes to use VTLN for creating synthetic speaker dependent models without using any data from the test speaker. The idea is to train multiple speaker independent (SI) models by scaling the training data with VTLN and selectively choosing one of the models that is acoustically close to the test speaker for performing recognition. We will show through recognition experiments that the proposed approach improves the performance of children speakers using models trained on adult speech.

The rest of the paper is organized as follows: First, we briefly discuss how speaker normalization is done using VTLN and speaker adaptive training (SAT). We then present our experimental setup and preliminary results. We will then discuss how to train synthetic speaker models using VTLN and show that the proposed approach performs better in mismatched task and speaker conditions.

2. Conventional Speaker Normalization

In this section, we briefly review VTLN and CMLLR and how they are combined to reduce speaker variability in ASR.

2.1. Vocal Tract Length Normalization (VTLN)

Vocal tract length normalization [5, 6] reduces speaker variability by frequency scaling the spectra of speech signals. Since it is difficult to choose a single speaker as reference for the entire population of speakers, a maximum likelihood (ML) search is followed in practice to find the optimal scaling factor and is given by:

$$\hat{\alpha}_{\mathbf{ML}} = \arg\max_{\alpha} p\{\mathbf{X}^{\alpha}|\lambda;\mathbf{W}\} \quad (1)$$

where, λ is the HMM model and $\mathbf{W}$ is the true transcription during training and first-pass transcription during recognition. $\mathbf{X}^{\alpha}$ are the VTLN-warped Mel-frequency cepstral coefficients (MFCC) appended with delta and acceleration coefficients. The range of α is between 0.80 and 1.20, with steps of 0.02. From Eq. 1, it is evident that the features are warped such that they move acoustically close to the model (λ). VTLN transforms the test speaker acoustically close to the trained model and can be seen as a way to derive *speaker independent features*.

2.2. Speaker Adaptive Training (SAT)

Speaker adaptive training (SAT) [14] applies CMLLR matrices [9, 10] to train speaker independent (SI) models. CMLLR is a constrained transformation approach to speaker adaptation that estimates a single matrix to transform both mean and covariance parameters of the model, i.e.

$$\hat{\mu} = \mathbf{B}\mu + b \quad \text{and} \quad \hat{\Sigma} = \mathbf{B}\Sigma\mathbf{B}^{T} \quad (2)$$

where, $\mathbf{B}$ is the CMLLR transformation matrix. μ and Σ are the mean and covariance parameters of the model respectively and b is the bias parameter. Each speaker in training or recognition has its associated matrix which is estimated by pooling all the data corresponding to that speaker.

CMLLR can also be seen as a feature transformation approach due to the constraint that the same matrix is used for transforming both means and covariances [10], i.e.

$$\mathcal{L}(\mathbf{X};\mu,\mathbf{\Sigma},\mathbf{B}) = \mathcal{L}(\mathbf{B^{-1}X};\mu,\mathbf{\Sigma}) + \log(|\mathbf{B^{-1}}|) \quad (3)$$

The above equation states that, the likelihood of the feature $\mathbf{X}$ given the model parameters mean (μ) and covariance (Σ) along with the transformation ($\mathbf{B}$) is equivalent to transforming the features with the inverse transformation and accounting for Jacobian. This means CMLLR performs feature transformation similar to VTLN, with the exception that it is learnt from the data. So CMLLR can also be seen as a way to derive *speaker independent features*.

2.3. Combining VTLN and SAT

VTLN and SAT are applied in combination to improve the performance of SI-ASR [12, 13] and is very common in large vocabulary continuous speech recognition (LVCSR). It is important to understand that the features need to be transformed with VTLN first before the SAT matrices are applied. This means, there is a restriction on the sequence of operations while combining VTLN and SAT. VTLN performs spectral scaling whereas SAT or CMLLR directly transform the MFCC features. So once the features are transformed using SAT, it is not known how to scale the spectra using VTLN and hence the restriction. Since VTLN and SAT are applied in our experiments to create SI models, we summarize the sequence of steps involved in training and recognition below.

2.3.1. Training

1. Create the baseline model (λ_B) using all the training data without any speaker normalization.
2. Create the VTLN model (λ_V) using the warped features estimated on the training set. The warp-factors are estimated using Eq. (1) w.r.t. λ_B and true transcription at utterance level.
3. Create the SAT model (λ_{VS}) by adapting the λ_V model obtained in Step 2. The CMLLR matrix is estimated using the VTLN-warped data corresponding to a particular speaker and λ_V. It is also possible to adapt the baseline model (λ_B) to create the SAT model (λ_{BS} - no VTLN), in which case the CMLLR matrix is estimated using the un-warped data and λ_B.

2.3.2. Recognition

1. Perform recognition using the baseline model (λ_B).
2. Perform recognition using the VTLN model (λ_V) and the VTLN-warped features estimated on the test set. The optimal α for the test speaker is estimated using λ_B and the transcription obtained in Step 1.
3. Perform recognition using λ_{VS}. The SAT matrices are estimated on VTLN-warped features, transcription obtained in Step 2 and λ_V. While performing recognition using λ_{BS}, the SAT matrices are estimated on un-warped data, transcription obtained in Step 1 and λ_B.

It is important to note that the same sequence of operations are applied both in training and recognition. Before proceeding further with our discussion, we present our experimental setup and preliminary results using VTLN and SAT.

3. Experimental Setup

All the experiments were done using HTK toolkit on WSJ0 [15], CMUKIDS [16] and TIDIGITS [17] corpus. Table 1 summarizes the different combinations of tasks investigated in our experiments. The WSJ0 task is a matched speaker experiment and includes adult male and female speakers in both training and recognition. TIDIGITS is a mismatched speaker experiment, where the models are trained using adult male speakers and are used for recognizing children speakers. CMUKIDS is a mismatched task and speaker experiment, where the models are trained on the TIMIT corpus [18] using both adult male and female speakers and are used for recognizing children speakers from the CMUKIDS database.

Table 1: *Summary of speech recognition tasks with train and test speaker conditions.*

Task	Train	Test
WSJ0	WSJ0 - Adults	WSJ0 - Adults
TIDIGITS	TIDIGITS - Adult Male	TIDIGITS - Children
CMUKIDS	TIMIT - Adults	CMUKIDS - Children

In the WSJ0 task, the models use context-dependent HMM's and have 3 emitting states with 16-component GMMs per state. We start with 41 context-independent phones that include silence and short-pause. In the TIDIGITS task, we use word models, that include *zero* to *nine* and *oh*. The digits were modeled with 16 emitting states with 5-component GMMs per state. Silence is modeled using a 3 state HMM having 6-component GMMs per state. For the CMUKIDS task, we train 39 context-independent HMM models, that have 3 emitting states with 32-component GMMs per state using the TIMIT corpus. The features in all the tasks are of 39 dimensions comprising MFCCs appended with delta and acceleration coefficients. In all cases cepstral mean subtraction is applied. A two-pass approach is followed in recognition for both VTLN and SAT.

For the WSJ0 task in recognition, we use the Nov. 92 test set, that has 8 speakers and has a total of 330 utterances. For the case of TIDIGITS, we have 50 children speakers and a total of 3847 utterances. The CMUKIDS database has two sets of recordings, namely SUM95 (the speakers were good readers) and FP (the speaker have more dialectical variations). For our experiments we use recordings from SUM95 set and only use utterances that belong to *bin1*, where the sentences were read correctly. This test set has 51 speakers and a total of 834 utterances.

3.1. Preliminary Results

Table 2 presents the preliminary results using VTLN and SAT to perform speaker normalization as stand alone operations and in combination. Please ignore the results for CMUKIDS in brackets for the time being. We will visit them in later sections of the paper. Observe that the baseline (λ_B) result without any transformation of the features or the model parameters has a huge degradation in performance for CMUKIDS and TIDIGITS. This degradation can be attributed to the difference in speaker characteristics between the train and test speakers. The performance improves with subsequent VTLN (λ_V) and SAT (λ_{BS}) as stand alone operations. The combined VTLN and SAT model (λ_{VS}) gives the best performance.

Note that the performance of λ_{BS} is inferior when compared with the performance of λ_V for CMUKIDS and TIDIGITS. A possible reason for this can be poor first-pass transcription for estimation of SAT matrices. On the other hand, CMLLR is a generic acoustic compensation technique where as VTLN is specifically designed to tackle the differences in VTL's and might be well suited for compensating the acoustic mismatch

Table 2: *Recognition performance (% WER) comparing VTLN and SAT for WSJ0, CMUKIDS and TIDIGITS tasks.*

	Model	WSJ0	CMUKIDS	TIDIGITS
Baseline (B)	λ_B	4.5	33.0	34.1
B + VTLN	λ_V	3.8	21.0 (20.5)	3.0
B + SAT	λ_{BS}	3.2	23.9 (22.8)	10.9
B + VTLN + SAT	λ_{VS}	2.6	17.8 (16.8)	1.6

between children and adults. The range of α is between 0.80 and 1.20 for WSJ0 task and between 0.60 and 1.20 for children in CMUKIDS and TIDIGITS tasks for VTLN in recognition.

4. Synthetic Speaker Models Using VTLN

The idea to train synthetic speaker models using VTLN draws inspiration from speaker adaptation, where the model parameters are updated or transformed to suit a particular test speaker. VTLN on the contrary transforms the features either in training or recognition to make them speaker independent. The question we are exploring: Can VTLN be used to move the training model acoustically close to the test speaker, thereby making it speaker dependent similar to speaker adaptation? Such an approach might be quite useful when there is very little or no data available from the test speaker to create speaker dependent models using speaker adaptation.

Since VTLN scaling can be used to transform the test speaker to move acoustically close to the training data, in principle it should also allow us to transform the training speaker to move acoustically close to the test speaker. If we can find the appropriate VTLN scaling to transform the training speaker acoustically close to the test speaker, the VTLN warped features of the training data can be used to create speaker models that are acoustically close to the test speaker.

The warp-factor estimation in VTLN (from Eq. (1)) performs a maximum likelihood search w.r.t to a model, that is built using the training data. Since we are proposing to estimate the scaling required to transform the training data rather the test data, the question will be, which model to use for obtaining the optimal α. Ideally, a model of the test data is required. Since it is not possible, we propose to train multiple VTLN warped models by scaling the training data with VTLN and selectively choosing one of the models that is acoustically close to the test speaker while performing recognition. The selected VTLN warped model for a particular test speaker is only a representative speaker model and does not use any information of the test speaker. Hence we call these VTLN warped models as *synthetic speaker models*.

The idea to train multiple VTLN warped models for performing VTLN in not new and has been previously proposed in literature. In [6, 19], it has been proposed to represent the distribution of each warp-factor with mixture of multivariate Gaussians. The optimal α for the test speaker is chosen for the distribution that yields the best likelihood. In [20], it has been used to compensate for Jacobian in VTLN. In [21], it has been explored to perform VTLN at phoneme level. The major difference between the proposed approach in this paper and previously proposed approaches is that, previously the multiple VTLN warped models were only used to estimate the optimal scaling factor in VTLN. We are proposing to use the VTLN warped models for performing recognition. Initial impression would be that both conventional VTLN and the proposed approach might have similar performance, which might be true for matched speaker conditions. We will show that the proposed approach is quite helpful in mismatched speaker conditions, especially for the case of children where we rely on models trained using adult speech.

The steps involved in training and recognition for the proposed approach are given below. VTLN and SAT are applied in training to make the models speaker independent. For our experiments, the range of α is between 0.80 and 1.20 with steps of 0.02 for WSJ0 (matched speaker) task, where as between 0.80 and 1.50 with steps of 0.02 for TIDIGITS and TIMIT (mismatched speaker) tasks during training.

4.1. Training

The following steps are followed for each value of α in the search range.

1. Warp all the available training data with a specific α. We denote this as $\texttt{Trn}^{\alpha}$.
2. Create the baseline model (λ_B^{α}) corresponding to a particular α using $\texttt{Trn}^{\alpha}$.
3. Create the VTLN model (λ_V^{α}) using warped-features of $\texttt{Trn}^{\alpha}$ for a particular α. This means VTLN is performed on top of already warped training data $\texttt{Trn}^{\alpha}$.
4. Create the SAT model (λ_{VS}^{α}) for a particular α by applying SAT matrices estimated using the VTLN warped features obtained in step 3.

By the end of training, we have a set of models corresponding to all the warp factors in the search range.

4.2. Recognition

1. Find the optimal warped model that is close to the test speaker using a maximum likelihood search and is given by:
$$\hat{\lambda}_{\mathbf{ML}}^{\alpha} = \arg\max_{\alpha} p\{\mathbf{X}|\lambda^{\alpha}; \mathbf{W}\} \quad (4)$$
The only difference compared with Eq. 1 is that, a set of warped models are used in place of a set of warped features. Note that the test features are un-warped. The warped-model that gives the best likelihood is chosen as the model that is acoustically close to the test speaker and is used for performing recognition. We select a single warped model for all the utterances corresponding to a particular test speaker.
2. VTLN in the proposed approach is implicitly included as a part of the training and it would not be necessary to apply VTLN in recognition. Since we choose a single warped model for all the utterances corresponding to a particular speaker, performing VTLN in recognition allows for variation of warped features at utterance level. For our experiments, the range of α during recognition is between 0.96 and 1.04 with steps of 0.02 only to allow small changes in spectral scaling.
3. The features of the test speaker can also be transformed using SAT matrices and the corresponding warped models are used for estimating the CMLLR matrices.

Table 3 presents the results using the proposed approach to reduce speaker variability. The results also include subsequent VTLN and SAT operations applied in recognition. We make the following observations in comparison with the results in Table 2:

Table 3: *Recognition performance (% WER) for WSJ0, CMUKIDS and TIDIGITS tasks using multiple VTLN-warped SI models with subsequent VTLN and SAT operations.*

	Model	WSJ0	CMUKIDS	TIDIGITS
Baseline (B)	λ_B^α	3.8	17.7	1.6
B + VTLN	λ_V^α	3.6	16.7	1.5
B + SAT	λ_{BS}^α	2.9	13.0	0.8
B + VTLN + SAT	λ_{VS}^α	2.4	11.8	0.8

- WSJ0 Task.
- The baseline result (λ_B^α) in the proposed approach is similar to performing VTLN (λ_V) in the conventional approach.
- The improvement observed with subsequent VTLN and SAT either as stand alone operations or in combination seem to provide small improvements.
- CMUKIDS and TIDIGITS Tasks.
- The baseline result (λ_B^α) already reaches the best performance that could be achieved using the conventional approach combining VTLN and SAT (λ_{VS}).
- Subsequent VTLN and SAT operations either as stand alone or in combination improve the performance of the proposed approach further. Comparing the best results in the conventional (λ_{VS}) and proposed (λ_{VS}^α) approaches, we see a relative improvement of 34% for CMUKIDS and 50% for TIDIGITS tasks.

The proposed approach seems to perform very well in mismatched speaker conditions.

5. Analysis and Discussion

The results corroborate the fact that VTLN can be used to create synthetic speaker models similar to speaker dependent models in speaker adaptation. Before we can conclusively say that the proposed approach helps improve the performance in mismatched speaker conditions, we address some more concerns. For analysis, we will be using the results from CMUKIDS task.

- Applying VTLN in the proposed approach during recognition is like applying another stage of VTLN on top of the warped features in the conventional approach, which inherently means a much finer grid search is performed and the proposed approach may be at an advantage.
- In order to understand this, we perform the VTLN experiment with a much finer grid in the conventional approach. The results are presented in Table. 2 for CMUKIDS task in brackets. We observe that there is small improvement in performance, but could not reach the performance of the proposed approach (Table. 3).
- Performing another stage of adaptation during recognition in the conventional approach with improved transcription might help us achieve a performance obtained in the proposed approach.
- The results are presented in Table. 2 for CMUKIDS in brackets. Applying SAT as a stand alone operation or when combined with VTLN seem to provide improvements in the conventional, but could not reach the performance of the proposed approach.

Table 4: *Recognition performance (% WER) for WSJ0, CMUKIDS and TIDIGITS tasks using the modified training procedure to bypass VTLN in training and apply only SAT.*

Train Model	Test Condition	WSJ0	CMUKIDS	TIDIGITS
λ_{BS}	VTLN + SAT	2.7	17.5	1.7
λ_{BS}^α	VTLN + SAT	2.5	11.7	0.7

These observations indicate that using synthetic speaker models might be better and seems to improve the performance in mismatched speaker conditions.

5.1. Simplifications to the proposed approach

The proposed approach comes with its own limitations in terms of implementation. Multiple SI models have to be trained in advance, with subsequent VTLN and SAT operations to obtain the VTLN warped SI models. This means a set of models have to be trained instead of a single model. Here we will discuss certain simplifications that can help reduce the computational load while implementing the proposed approach.

- Training multiple VTLN warped models can be simplified by representing VTLN scaling as a matrix operation [22]. Considering the VTLN matrix as a CMLLR transformation common for all the speakers in the training data, we can easily adapt λ_B to λ_B^α.
- If VTLN and SAT are applied while training the SI model, we will show through recognition experiments that it would be sufficient to perform only SAT and bypass VTLN in training. VTLN is applied as usual in recognition.

  Table 4 presents the results using the proposed modification in training, where only SAT is applied in training bypassing the VTLN normalization. Comparing the result of λ_{BS} in Table 4 with λ_{VS} in Table 2 and λ_{BS}^α in Table 4 with λ_{VS}^α in Table 3, we observe that the performance is almost similar. The results indicate that VTLN can be bypassed in training and it is sufficient only to apply SAT. The transcription of VTLN warped features, required for estimating the SAT matrices in recognition is generated by performing recognition using the SAT model instead of the VTLN model.

The above discussed simplification make the training procedure fast, but this does not mean that they are necessary for training synthetic speaker models.

6. Conclusion

In this paper, we proposed an approach to train synthetic speaker models using VTLN. The approach draws inspiration from speaker adaptation based approaches, that transform or update the model parameters of the SI model to make it speaker dependent. We proposed to train multiple VTLN-warped SI models by scaling the training data with VTLN and choosing one of the models that is acoustically close to the test speaker while performing recognition. We believe such an approach might be quite helpful when there is very little or no data available from the test speaker to create speaker dependent models using speaker adaptation. We showed that the proposed approach improves the performance in mismatched speaker conditions, especially while recognizing children speakers using models trained on adult speech.

7. References

[1] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," *Proc. of ICASSP*, Hong Kong, China, Apr. 2003.

[2] A. Potamianos, S. Narayanan, and S. Lee, "Automatic Speech Recognition for Children," in *Proc. of EUROSPEECH*, Rhodes, Greece, Sept. 1997.

[3] J. G. Wilpon and C. N. Jacobsen, "A Study of Speech Recognition for Children and Elderly," in *Proc. of ICASSP*, Atlanta, GA, May 1996.

[4] S. Das, D. Nix, and M. Picheny, "Improvements in Childrens Speech Recognition Performance," in *Proc. of ICASSP*, Seattle, WA, May 1998.

[5] A. Andreou, T. Kamm, and J. Cohen, "Experiments in Vocal Tract Normalization," in *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[6] L. Lee and R. Rose, "Frequency Warping Approach to Speaker Normalization," *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 49–59, Jan. 1998.

[7] J. Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291-298, Apr 1994.

[8] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171 – 185, 1995.

[9] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker Adaptation using Constrained Estimation of Gaussian Mixtures," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 5, pp. 357 –366, Sep. 1995.

[10] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[11] M. Gerosa, D. Giuliani and F. Brugnara, "Acoustic variability and automatic recognition of childrens speech", *Speech Communication*, Vol. 49, pp. 847-860, Oct.-Nov. 2007.

[12] D. Giuliani, M. Gerosa, and F. Brugnara, "Speaker Normalization through Constrained MLLR Based Transforms," in *Proc. of Interspeech*, Jeju Island, Korea, Oct. 2004.

[13] D. R. Sanand and M. Kurimo, "A Study on Combining VTLN and SAT to Improve the Performance of Automatic Speech Recognition," in *Proc. of Interspeech*, Florence, Italy, Aug. 2011.

[14] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," in *Proc. of ICSLP*, pp. 1137–1140, Oct. 1996.

[15] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-Based CSR Corpus," in *Proc. of ICSLP*, Banff, Alberta, Canada, Oct. 1992.

[16] M. Eskenazi, J. Mostow, and D. Graff, "The CMU Kids Corpus." Linguistic Data Consortium, 1997.

[17] R. Leonard, "A Database for Speaker-Independent Digit Recognition," in *Proc. of ICASSP*, San Diego, California, USA, Mar. 1984.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S.Pallett, N. L. Dahlgrena, and V. Zue, "Timit Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, 1993.

[19] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," in *Proc. ICASSP*, Phoenix, AZ, Mar. 1999.

[20] R. Sinha and S. Umesh, "A Method for Compensation of Jacobian in Speaker Normalization," in *Proc. of ICASSP*, vol. 1, pp. 560–563, Apr. 2003.

[21] M. Blomberg and D. Elenius, "Vocal Tract Length Compnesation in Signal and Model Domains in Child Speech Recognition," in *Proc. Fonetik*, vol. 50, no. 1, pp. 41–44, 2007.

[22] D. R. Sanand and S. Umesh, "VTLN Using Analytically Determined Linear-Transformation on Conventional MFCC," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, pp. 1573–1584, Jul. 2012.