Text-Independent Speaker Identification Using Vocal Tract Length Normalization for Building Universal Background Model

A. K. Sarkar, S. Umesh and S. P. Rath

Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur, India-208016

[achintya,sumesh,srath]@iitk.ac.in

Abstract

In this paper, we propose to use Vocal Tract Length Normalization (VTLN) to build the Universal Background Model (UBM) for a closed set speaker identification system. Vocal Tract Length (VTL) differences among speakers is a major source of variability in the speech signal. Since the UBM model is trained using data from many speakers, it statistically captures this inherent variation in the speech signal, which results in a "coarse" model in the acoustic space. This may cause the adapted speaker models obtained from the UBM model to have significantly high overlap in the acoustic space. We hypothesize that the use of VTLN will help in compacting the UBM model and thus the speaker adapted models obtained from this compact model will have better speaker-separability in the acoustic space. We perform experiments on MIT, TIMIT and NIST 2004 SRE databases and show that using VTLN we can achieve lesser Identification Error Rates as compared to the conventional GMM-UBM based method.

Index Terms: Speaker Identification, VTLN, Iterative MAP, UBM

1. Introduction

In Automatic Speech Recognition (ASR), the Speaker Independent (SI) HMM model is built using data from a large number of speakers in the training set. This enables the model to capture the speaker-variations in the speech signal, and therefore, works reasonably well for any arbitrary speaker in the test set. Because of these variations in the spectra, the SI model is coarser in the acoustic space when compared to a Speaker Dependent (SD) model trained using data taken from a particular speaker. This, in turn, results in poorer word recognition performance of SI systems in comparison with SD systems for that speaker.

Vocal Tract Length Normalization (VTLN) is a commonly used technique that reduces this variabilities in the speech signal by frequency-warping the spectra of a particular speaker. The frequency-warped utterances are then used to build the Speaker-Normalized VTLN model. It has been observed that the word recognition performance of these Normalized-VTLN models is better than Un-normalized-SI models and approach that of SD models. This is an indication of the fact that, the VTLN-normalized model is a more compact model than the unnormalized model.

Analogous to SI-ASR, a Universal Background Model (UBM) is built for Speaker Identification using training data from many different speakers in the training set. We hypothesize that this results in a *coarse* GMM-UBM model due to inherent inter-speaker variability in the speech signal. In this paper, we propose to use VTLN to build a *compact* Universal Background Model (UBM) for Speaker Identification so that after adaptation the target speaker models are better separated.

In sec. 2 we discuss the motivation behind using VTLN to build a Universal Background Model. In sec. 3 we describe the procedure for building VTLN-UBM followed by a description of the experimental setup in sec. 4, likelihood analysis in sec. 5 and experimental results in sec. 6. Sec. 7 concludes the paper.

2. Motivation for Using VTLN

As mentioned before, the UBM model is trained using data of all the speakers in the training sets leading to a coarse UBM model due to spectral variations. Adaptation of this coarse model may result in the target speaker models having a significant amount of overlap in the acoustic space. We propose the use of VTLN to minimize spectral variations among speech utterances and build a VTLN-UBM model using the VTLN-warped (i.e. normalized) utterances. Target speaker models are then obtained by adapting this canonical VTLN-UBM model with the help of adaptation data. Adaptation of VTLN-UBM would result in a compact model for the target speaker, much like what could have been built if sufficient data from that speaker were available. This should reduce the overlap among speaker models and therefore yield better speaker identification accuracy. This is illustrated in Fig. 1.

Our proposed method has some similarity with the method used in [1] to build a GMM-UBM by repeatedly estimating CM-LLR matrices for background speakers and building a canonical model. However, the subsequent modeling in [1] is done using support-vector machines (SVMs). In this paper, we use cepstral feature based systems to analyze the performance of VTLN-UBM and compare it with the conventional GMM-UBM.

In the next section, we describe the procedure used to build VTLN-UBM.

3. Building VTLN-UBM Model

VTLN reduces inter-speaker variability by warping the spectrum of one speaker so as to match the spectrum of another speaker [2]. Since in practice there is no reference speaker with respect to whom we can find the scaling (or warping) factor, an ML based approach is followed to estimate the frequency-warp-factor, α. The ML estimate is obtained by doing a grid-search for α over the range [0.8,1.20] based on physiological constraints on the vocal-tract. We have recently proposed efficient [3] and faster [4] alternatives for conventional VTLN.

In this paper, the following steps are followed to build VTLN-UBM model:

Initial Step: Generate and store the VTLN-warped utterances, $\{X_r^\alpha\}_{r=1}^R$ (X_r denotes the r^{th} utterance), for all speech utterances in the UBM training set over a set of values of α. Set the GMM-UBM as the λ_0 model, $i = 0$ and repeat the follow-

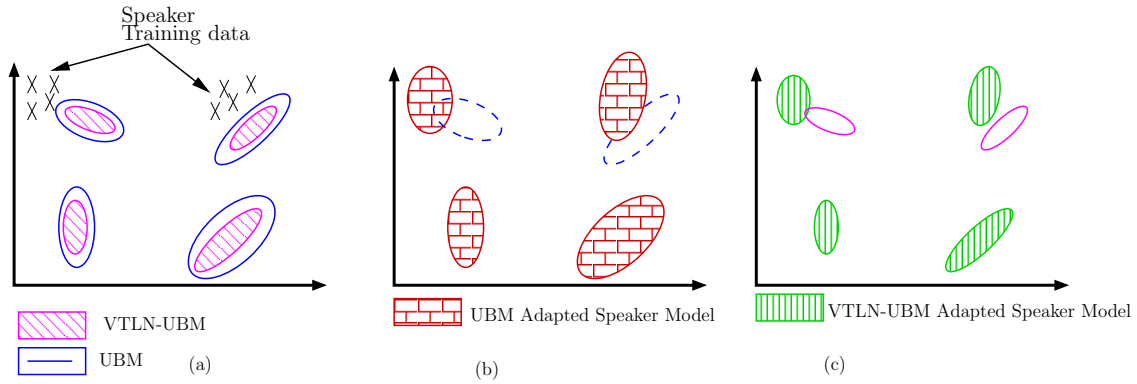

Figure 1: *Illustration of GMM-UBM and VTLN-UBM adaptation: (a) VTLN-UBM is more compact (i.e. small in space) than GMM-UBM. (b,c) Speaker adapted models from VTLN-UBM have better speaker class separability than GMM-UBM adapted models.*

ing steps a few times.

1. Estimate the warp-factors for the training utterances by maximizing the likelihood of the warped utterances w.r.t. the λ_i model, i.e.

$$\hat{\alpha}_r = \arg \max_{\alpha} Pr(X_r^{\alpha}|\lambda_i) \quad (1)$$

2. Build λ_{i+1} VTLN-UBM (Normalized) model using warped utterances, i.e.,

$$\lambda_{i+1} = \arg \max_{\lambda} Pr(\{X_r^{\hat{\alpha}_r}\}_{r=1}^{R}|\lambda) \quad (2)$$

3. Set $i = i + 1$ and go to step 1.

We now present a series of experiments to understand the role of VTLN in compacting the UBM and providing better separability of adapted speaker models.

4. Experimental Setup

All the experiments were performed on MIT, TIMIT and NIST 2004 SRE corpora using HTK[5] and ALIZE toolkit[6]. TIMIT contains 10 phonetically balanced utterances spoken by each of the 630 speakers in the database. Data from 462 speakers (including 326 males and 136 females) were used for adaptation and testing. Eight utterances from each speaker were used to adapt the UBM and the other 2 utterances were used for testing [7]. All the 10 utterances from the remaining 168 speakers were used to build the UBM model.

In MIT corpus [8], 48 speakers (26 males and 22 females) in the Enrollment Session were used for adaptation and testing. Speaker adapted models were obtained from the UBM by taking 54 adaptation utterances per speaker from Enrollment Session 1. For testing, 54 utterances per speaker from Enrollment Session 2 were used. The UBM model was built using all the speakers from the imposter directory.

Using NIST 2004 SRE, the experiment was performed on 1 side train-10 second test data as per the evaluation plan in [9]. The UBM was trained using data from NIST 2002 SRE(cell & landline) and Switchboard-1 Release-2(landline) databases. Data from 310 unique speakers in 1 side conversation were used to build speaker models by adapting the UBM. The test data consists of 1163 utterances from 306 speakers.

In all our experiments, we use 39 dimensional MFCC feature vectors(c_1 to c_{13} with their Δ and Δ-Δ coefficients, excluding c_0). In order to extract features, filter bank coefficients are first computed over 20ms Hamming windowed frames at a frame rate of 10ms. For MIT and TIMIT databases, no front end signal pre-processing was done during feature extraction. After feature extraction, cepstral mean subtraction(utterance wise) was applied on MIT data. It was not done on TIMIT data because there is no significant variation in the microphone used for recording or in the channel. For the NIST 2004 SRE system, we use two different frame removal techniques as in [10]. One is a bi-gaussian modeling of energy components(for NIST 2002 SRE & Switchboard-1 Release-2 corpora) and the other is a tri-gaussian modeling of 0-mean and 1-variance normalized energy components (for NIST 2004 SRE). Finally, silence removed features are normalized to fit a 0-mean and 1-variance distribution(utterance wise).

The UBM in all databases contains 2048 gaussians with diagonal co-variance matrices except in the NIST 2004 corpus where we consider a smaller UBM model in order to further investigate our proposed method. Only the means of the UBM mixture components are adapted using Maximum a-posteriori (MAP) technique to build speaker models. During identification, we use the likelihood computed from all the components.

5. Analysis of GMM-UBM and VTLN-UBM

In order to understand the difference between GMM-UBM and VTLN-UBM, we perform three different types of analysis. In all of them, the average likelihood for a speaker is obtained by averaging likelihoods of all the adaptation utterances from that particular speaker. To get a better understanding, speaker indices in the figures are sorted in order of increasing likelihood.

5.1. Likelihood Before and After Adaptation

The average likelihood before and after adaptation with respect to both GMM-UBM and VTLN-UBM systems is plotted in Fig.2 and 3 for TIMIT and MIT databases respectively and in Fig.4 (which shows in terms of differences in likelihood) for NIST 2004 SRE. From the figures, we can observe that:

• The average likelihood of adaptation utterances with respect to VTLN-UBM are lower for most of the speakers when compared to GMM-UBM. This is as expected, since the un-warped adaptation utterances would often be further away from the compact VTLN-UBM. An analogous observation also occurs in speech recognition, where the performance of *un-warped* utterances with VTLN-HMM is often poorer than with SI model.

• The average likelihood of adaptation utterances with respect to the adapted speaker model obtained from VTLN-UBM is

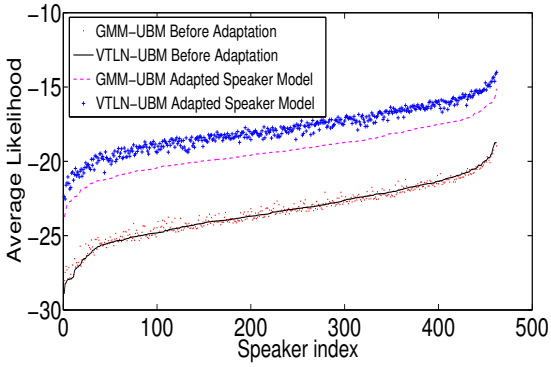

Figure 2: *Speaker-wise average likelihood of adaptation data in TIMIT. Note that* after adaptation *VTLN-UBM likelihood is higher than that of GMM-UBM*

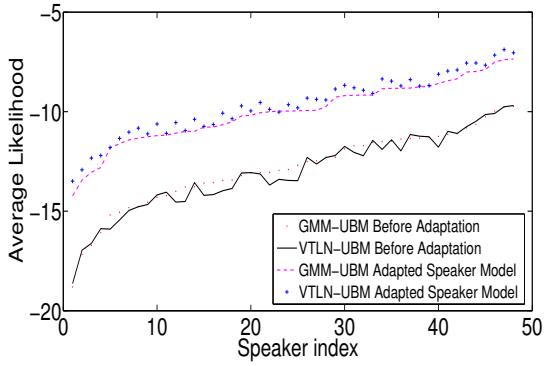

Figure 3: *Speaker-wise average likelihood of adaptation data in MIT.*

always *better* than the likelihood with respect to the adapted speaker model obtained from GMM-UBM indicating a better adapted model.

5.2. Multiple Iterations of MAP

We now investigate the gain in likelihood using multiple iterations of MAP. In Fig.4 and 5, the average likelihood ratios $\Lambda_{GMM}(X)$ and $\Lambda_{VTLN}(X)$ are plotted for NIST 2004 SRE. $\Lambda_{GMM}(X) = log\ p(X|\lambda^{'}_{GMM}) - log\ p(X|\lambda_{GMM})$ where X is the feature vector, $\lambda^{'}_{GMM}$ is the model adapted from GMM-UBM and λ_{GMM} is the GMM-UBM. Similarly, $\Lambda_{VTLN}(X) = log\ p(X|\lambda^{'}_{VTLN}) - log\ p(X|\lambda_{VTLN})$ where $\lambda^{'}_{VTLN}$ is the model adapted from VTLN-UBM and λ_{VTLN} is the VTLN-UBM.

• Fig.4 shows average $\Lambda_{GMM}(X)$ and $\Lambda_{VTLN}(X)$ for 128 component mixture models after the first and second iterations of MAP adaptation. It can be observed that in the first iteration of MAP, $\Lambda_{VTLN}(X)$ is more than $\Lambda_{GMM}(X)$ for majority of the speakers. This indicates that for most of the speakers, $\lambda^{'}_{VTLN}$ is better adapted than $\lambda^{'}_{GMM}$ after the first iteration. This fact is also supported by a slight IER reduction, as seen in table 1. After the second iteration of MAP, it can be observed that more than 80% of the speakers have $\Lambda_{VTLN}(X)$ greater than $\Lambda_{GMM}(X)$. Hence the second iteration provides significant improvement in adaptation over the first for the VTLN case. This is also clear from a significant IER reduction after the second iteration, as seen in table 1.

• Fig.5 shows a similar analysis for the 512 component mixture models. In this case after the first iteration of MAP,

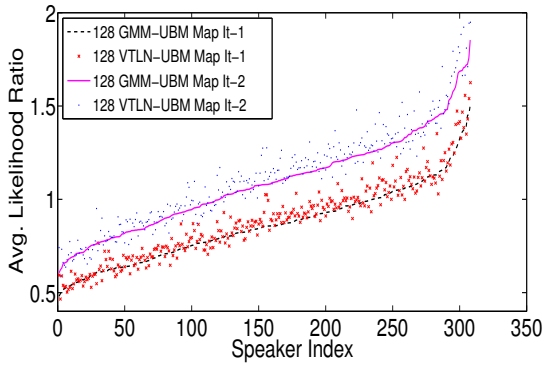

Figure 4: Λ_{VTLN} *and* Λ_{GMM} *after the first and second iteration of MAP for a 128 mixture model built using NIST 2004 data. It shows that adaptation is better after the second MAP iteration.*

$\Lambda_{VTLN}(X)$ is seen to be lesser than $\Lambda_{GMM}(X)$ for majority of the speakers. This indicates that the first iteration does not adapt the VTLN-UBM models very well. The same is indicated by a poor IER performance as seen in table 1. However, after the second iteration of MAP, more than half of the speakers have a $\Lambda_{VTLN}(X)$ which is higher than $\Lambda_{GMM}(X)$. Again the second iteration performs better than the first, which can also be seen from the IER performance in table 1.

We can deduce from the above observations that since VTLN-UBM is compact, the adaptation data and the model lie far apart in the acoustic space. Therefore, VTLN-UBM does not move significantly after the first MAP iteration. More number of iterations are, therefore, required for better adaptation.

5.3. Comparison of Competing Model's Likelihood

We analyze the likelihood of the GMM-UBM and VTLN-UBM in terms of separability of the speaker-adapted model. We hypothesize that since the VTLN-UBM-adapted model is more compact than the GMM-UBM-adapted model, the likelihood of the competing model when compared to the correct(true speaker) model should be significantly lower. In Fig.6, the likelihood ratio $\Lambda_{NN}(X)$ is plotted for NIST 2004 SRE. $\Lambda_{NN}(X) = log\ p(X|\lambda^{'}_{TRUE}) - log\ p(X|\lambda^{'}_{CMP})$ where $\lambda^{'}_{TRUE}$ is the true speaker's adapted model and $\lambda^{'}_{CMP}$ is the adapted model of the *nearest* competing neighbor. In order to find the nearest neighbor, likelihoods of the true speaker data are

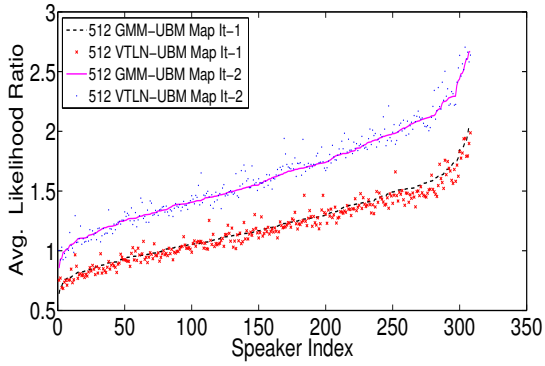

Figure 5: Λ_{VTLN} *and* Λ_{GMM} *after the first and second iteration of MAP for a 512 mixture model built using NIST 2004 data. It shows that adaptation is better after the second MAP iteration.*

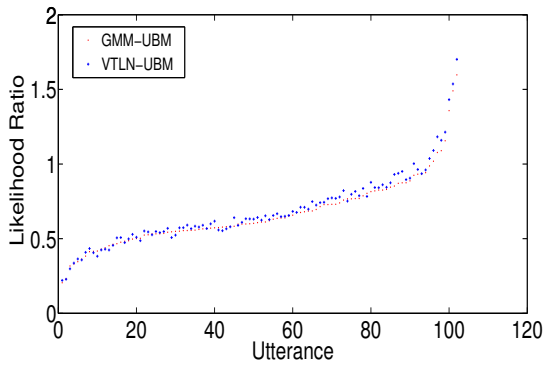

Figure 6: $\Lambda_{NN}(X)$ *plotted for VTLN-UBM and GMM-UBM using NIST 2004 data. It depicts the better separability of VTLN-UBM adapted models.*

computed from all models apart from the true speaker model, and the one that yields the maximum value is chosen.

• Fig. 6 shows that the likelihood of the true speaker model is much *higher* than the nearest competing speaker model for VTLN-UBM system when compared to GMM-UBM system. This indicates better separability of VTLN-UBM adapted speaker models validating our conjecture that the VTLN-UBM-adapted models are more compact.

6. Identification Results and Discussion

We now compare the GMM-UBM and VTLN-UBM systems in terms of Identification Error Rates (IER):

• From Table 1, it is seen that there is a slight IER reduction using VTLN-UBM over standard UBM for all three databases namely, MIT, TIMIT and NIST 2004 SRE. The relevance factor (τ) in MAP adaptation is indicated within parenthesis. Since GMM-UBM and VTLN-UBM are independent models, the value of τ need not be the same when adapted models are obtained from these UBMs.

• Table 1 also shows the IER for multiple iterations of MAP, and it can be seen that there is significant improvement in VTLN-UBM performance in the second iteration especially for NIST 2004 SRE .

• Table 2 shows the N-Best IER results for NIST 2004 SRE(for 128 components), MIT and TIMIT databases. We see that VTLN-UBM performs almost consistently better than GMM-UBM for all N.

Table 1: Identification Error Rate (%) Performance on NIST 2004 SRE, MIT and TIMIT databases

Corpus	UBM size	System	# of MAP Iteration	
			1	2
NIST 2004	128	GMM-UBM	65.35 (6)	65.69 (10)
		VTLN-UBM	65.18 (6)	63.54 (10)
	256	GMM-UBM	62.5 (8)	62.42 (10)
		VTLN-UBM	62.94 (6)	62.34 (10)
	512	GMM-UBM	61.05 (6)	60.96 (12)
		VTLN-UBM	61.13 (6)	60.79 (6)
MIT	2048	GMM-UBM	7.56 (7)	7.37 (8)
		VTLN-UBM	7.21 (5)	7.18 (34)
TIMIT	2048	GMM-UBM	0.87 (7)	no change
		VTLN-UBM	0.65 (4)	no change

Table 2: Identification Error Rate (%) Performance for N-Best

Corpus	System	N-Best			
		2	3	4	5
NIST 2004	GMM-UBM	56.32	51.25	47.21	44.54
	VTLN-UBM	55.28	50.47	47.03	44.37
MIT	GMM-UBM	4.05	2.31	1.70	1.47
	VTLN-UBM	3.82	1.93	1.39	1.16
TIMIT	GMM-UBM	0.11	0.11	0	0
	VTLN-UBM	0.32	0.11	0	0

7. Conclusions

In this paper, we have proposed the use of VTLN in building a UBM model for speaker identification. The motivation for our proposed approach is to obtain more compact speaker dependent models after adaptation. This is because VTLN reduces inter-speaker variability caused by differences in VTL among speakers and hence the VTLN-UBM model has lesser such variabilities. We have shown that the speaker-adapted model obtained from VTLN-UBM provides higher likelihood than the GMM-UBM adapted model indicating that the VTLN-UBM-adapted models are better. We have also observed that VTLN-UBM has larger likelihood difference between the correct and the nearest competing model when compared to conventional GMM-UBM and this is reflected in the lower identification error rate for VTLN-UBM. Finally, we have shown that using multiple iterations of MAP helps the VTLN-UBM system significantly.

8. Acknowledgement

A part of this work was supported by SERC project funding `SR/S3/EECE/058/2008` from the Department of Science & Technology, Ministry of Science & Technology, India.

9. References

[1] M. Ferras, C. C. Leung, C. Barras, and J. L. Gauvain, "Constrained MLLR for Speaker Recognition," *ICASSP-2007*, vol. 4, pp. 53–56, 2007.

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

[3] D. R. Sanand and S. Umesh, "Study of Jacobian Compensation Using Linear Transformation of Conventional MFCC for VTLN," in *Interspeech2008*, September 2008.

[4] P. T. Akhil, S. P. Rath, S. Umesh, and D. R. sanand, "A Computationally Efficient Approach to Warp Factor Estimation in VTLN Using EM Algorithm and Suffi cient Statistics," in *Interspeech2008*, September 2008.

[5] S. Young, D. Kershaw, J. Odell, V. Valtchev, P. Woodland, and et al., "HTK Book," *Copyright 2001-2006 CUED*.

[6] "http://lia.univ-avignon.fr/heberges/alize/."

[7] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[8] R. Woo, A. Park, and T. J. Hazen, "The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments," *in Proc. IEEE Odyssey 2006-The speaker and Language Rec. Workshop, San Jaun, Puerto Rico*, pp. 1–6, 2006.

[9] The Evaluation Plan of NIST 2004 Speaker Recognition Campaign. http://www.itl.nist.gov/iad/mig//tests/sre/2004/SRE04_evalplan-v1a.pdf.

[10] J. F. Bonastre, N. Scheffer, C. Fredouille, and D. Matrouf, "Nist'04 Speaker Recognition Evaluation Campaign: New LIA Speaker Detection Plateform based on ALIZE Toolkit," *in NIST SRE'04 Workshop, Toledo, Spain*, Jun. 2004.