Efficient Speaker and Noise Normalization for Robust Speech Recognition

Vikas Joshi, Raghavendra Bilgi, S. Umesh

Department of Electrical Engineering
Indian Institute of Technology, Madras, India
ee10s001, ee10s009, umeshs@ee.iitm.ac.in

C. Benitez, L. Garcia

Dept of Signal Theory, Telematics and Communications
University of Granada, Spain
carmen, luzgm@urg.es

Abstract

In this paper, we describe a computationally efficient approach for combining speaker and noise normalization techniques. In particular, we combine the simple yet effective Histogram Equalization (HEQ) for noise compensation with Vocal-tract length normalization (VTLN) for speaker-normalization. While it is intuitive to remove noise first and then perform VTLN, this is difficult since HEQ performs noise compensation in the cepstral domain, while VTLN involves warping in spectral domain. In this paper, we investigate the use of the recently proposed T-VTLN approach to speaker normalization where matrix transformations are directly applied on cepstral features. We show that the speaker-specific warp-factors estimated even from noisy speech using this approach closely match those from clean-speech. Further, using sub-band HEQ (S-HEQ) and T-VTLN we get a significant relative improvement of 20% and an impressive 33.54% over baseline in recognition accuracy for Aurora-2 and Aurora-4 task respectively.

Index Terms: VTLN, T-VTLN, Robust features, Noise Compensation, HEQ, Sub-band HEQ

1. Introduction

The performance of a Speech recognition system degrades significantly when there is a mismatch in train and test environment. Noise and Inter-Speaker variability are two major sources of mismatch. One of the main reasons for inter-speaker variability is due to physiological differences in vocal tracts of speakers, which results in differences in spectra for same sound enunciated by different speakers. Vocal Tract Length Normalization (VTLN) [1] is a standard technique which compensates for this variability. VTLN operates in the filter bank domain where the spectra are compressed or expanded such that the corresponding features best match the model. One of the simplest and effective noise compensation technique has been Histogram Equalization (HEQ)[2] which makes no assumption about noise or speech characteristics.

VTLN and Noise compensation techniques have evolved separately. While noise compensation studies have only focussed on improving the performance of noisy speech without worrying about inter-speaker variability, most VTLN studies have focussed on reducing inter-speaker variability and have not investigated the effects of noise on VTLN. As we will show, VTLN performs even worse than baseline in the low SNR conditions. In VTLN, the scaling factor is found in a maximum-likelihood (ML) framework using Eqn. 1. This is implemented as a grid search by varying alpha (scaling or warp factor) from 0.8 to 1.2 with an increment of 0.02. The range of alpha is fixed based on physiological constraints. Therefore,

$$\hat{\alpha} = \arg\max_{\alpha} p(X^{\alpha}/\lambda, U) \qquad (1)$$

where X^{α} is warped cepstra, λ is the reference model, U is the transcription (first pass transcription is used in test cases). VTLN, however, is susceptible to noisy features and errors in first pass transcription which lead to alignment errors. Hence, performance degrades with noise. To make VTLN more robust against noise and to improve the recognition performance, noise compensation techniques need to be combined with VTLN.

We have recently discussed in [3] one technique, where VTLN and Histogram Equalization (HEQ) technique are combined to jointly compensate the effect of noise and speaker variability. This approach uses conventional VTLN technique, where features are first warped (in Filter bank domain) and then equalized to get a new set of features. While such an approach gives improved performance, this method has inherent problems associated with it, namely

- As VTLN is performed in filter bank (FB) domain before the noise compensation, noise spectra is also warped.
- 21 set of equalizations (corresponding to 21 warp-factors) are needed for each cepstral coefficient making it computationally more cumbersome.

This approach becomes even more cumbersome, when combined with our recently proposed sub-band Histogram Equalization (S-HEQ)[4] technique which requires additional two histogram equalizations.

A more intuitive and elegant approach would be to first eliminate the noise and then perform VTLN. This strategy is not possible in our previous approach as conventional VTLN is performed in filter-bank (FB) domain, while HEQ is done in cepstral domain. Applying HEQ itself in FB domain is not appropriate as there is a strong correlation between the FB energies. Further, the number of equalizations needed would be large. In this paper, we present a modified approach by exploiting the advantages of Linear transformation approach to VTLN (LT-VTLN). There have been several approaches proposed in literature to perform VTLN using linear transformations [5, 6, 7, 8]. We use the approach described in our earlier work [8, 9], referred to as Transform-VTLN (T-VTLN) since it can be applied on conventional MFCC unlike some of the other methods. T-VTLN has been shown to be computationally efficient and can be *directly* applied in cepstral domain, making noise compensation *before* VTLN-warping feasible. The usage of T-VTLN instead of conventional VTLN also makes it computationally more efficient in terms of warp-factor estimation. Moreover, only one HEQ transformation per cepstra is needed compared to 21 in our previous approach [3]. As we will show later, the match between VTLN warp factors from clean-speech and noisy speech, is better for our proposed approach compared to method in [3]. Recognition results also show comparable performance for the previous and proposed approach.

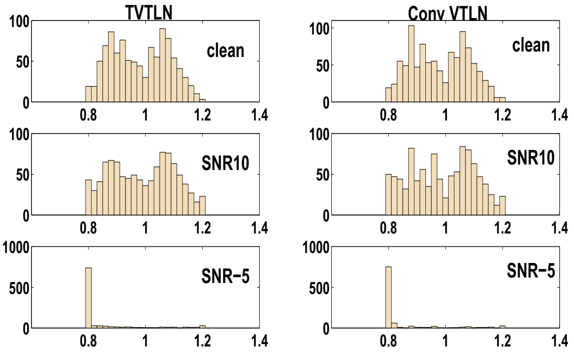


Figure 1: Warp-factor distribution for different SNR levels for T-VTLN and conventional VTLN using test utterances from Aurora-2

The advantage of the proposed approach is that it is computationally efficient and better matches the clean-speech VTLN warp-factors. Finally, T-VTLN becomes even more important, if we use Sub-band HEQ (S-HEQ), an improved approach to equalization, proposed in a companion paper [4] which involves *three* equalizations. Combing S-HEQ and T-VTLN provides a significant relative improvement of 33.54% in the recognition accuracy over baseline system for Aurora-4 database.

The paper is organised as follows. Section 2 contains a study of T-VTLN in noisy conditions. We then investigate the combined approaches of histogram-equalization and VTLN in Section. 3. S-HEQ approach is briefly explained in Section. 4, followed by experimental results and discussion in Section. 5. Finally, conclusions are presented in Section. 6.

2. T-VTLN in noisy conditions

In T-VTLN, the warped features are generated by using linear transformation of cepstra, i.e.

$$C^{\alpha} = T^{\alpha} * C^{1.0} \tag{2}$$

where C^{α} are the warped cepstra, T^{α} is Linear Transformation (LT) Matrix for warp-factor α and $C^{1.0}$ unwarped cepstra (i.e. conventional MFCC). LT approach leads to exact representation of the conventional VTLN as long as the cepstral coefficients are *quefrency limited* [8]. Conventional MFCC uses only 23 filter banks which may not result in completely que-frency limited cepstra. In order to match the T-VTLN cepstra with conventional VTLN, the following modifications were done in the conventional signal processing steps. (i) Additional half filters were added at the zero and Nyquist/2 frequency. (ii) 40 filters were used instead of 23, with equivalent bandwidth of using 12 filters (to make it quefrency limited). (iii) 21 cepstra are used for the transformation to get warped cepstra (to overcome truncation errors). For each warp-factor a $21*13$ T-VTLN matrix is generated, that transforms 21 dimensional unwarped cepstra to 13 dimensional warped cepstra. Transformed 13 cepstra were used for training and testing. These modifications are used *only* to ensure exact match between conventional and T-VTLN. As we will show later, T-VTLN performs equally well using conventional filter-bank and cepstra. After modification, warp-factor distributions for clean test utterances of Aurora-2 database have a correlation coefficient of 0.972, indicating a good match between conventional VTLN and T-VTLN. Similarly, recognition results also match for T-VTLN and conventional VTLN as shown in Tables 1 and 2.

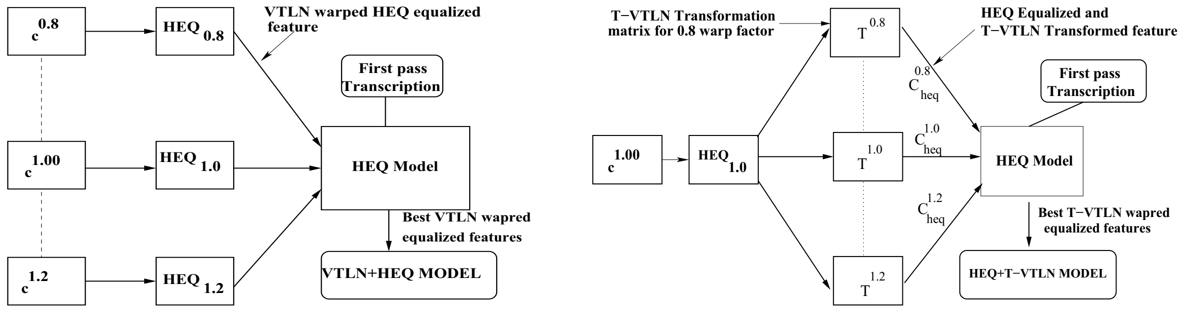
VTLN is sensitive to noise in features and errors in first pass transcription. With SNR degradation, first pass transcriptions generated by baseline system would be erroneous, resulting in alignment errors while estimating best warp factors. Presence of noise in features also affect the warp-factor estimation. The warp-factor distributions show aggressive warping of 0.8 at low SNR levels as shown in Fig. 1. Such a behaviour is due to increase in noise energy in high frequency regions, while speech has most of its energy concentrated in low frequency regions. Due to more energy in higher frequency region VTLN attempts to compress the spectra irrespective of speaker characteristics. The warp-factor distribution presented here do not exactly match the distributions presented in previous work [3] due to modifications in signal processing steps. As seen from the figures, VTLN warp-factor estimation completely breaks down in low-SNR conditions, and this is also reflected in the performance being even below that of baseline at low-SNR. In the next section, we describe methods to improve VTLN performance even in low-SNR conditions by combining with HEQ.

3. Combined Approach : VTLN+HEQ and HEQ+TVTLN

We refer to our previous approach [3] to combine VTLN and HEQ as VTLN+HEQ and proposed approach as HEQ+TVTLN. Fig. 2(a) shows implementation of VTLN+HEQ, where the features are first VTLN warped and then equalized. 21 set of warped features are obtained by warping, and corresponding 21 ref CDF's (for each cepstra) are generated. However in [3], *only one CDF (HEQ 1.0, CDF for α = 1.0 features)* was used for equalization for all the warped features. The best warp factor is estimated using VTLN+HEQ transformed features. HEQ model is used for alignment in training. The features corresponding to best α are chosen to build a speaker noise normalized model (termed as VTLN+HEQ model). This approach improves the performance over VTLN and HEQ alone. The sensitivity of warp-factors to SNR degradation is reduced compared to VTLN alone. However, this method has its inherent problems as discussed in the Introduction (Section 1).

A more elegant and a computationally efficient approach for speaker and noise normalization would be to use LT approach to VTLN (i.e T-VTLN), which allows us to first equalize the features and then perform the speaker normalization. Fig. 2(b) explains this approach. The training and the test strategy remains same as that of VTLN+HEQ as discussed in [3]. However, the steps for feature generation vary. First, equalization is performed on the unwarped features using the ref. CDF (obtained using the clean training utterances). Then warping is done by transforming equalized features using T-VTLN matrices which are analytically calculated and stored. The best-warp factor and corresponding HEQ+TVTLN features are estimated using the HEQ model. These features are used to build HEQ+TVTLN Model. Testing follows a similar approach, where best HEQ+TVTLN warped features are obtained using HEQ+TVTLN trained model. These optimally warped HEQ+TVTLN features are then used for final recognition.

We now analyze the warp-factor distribution obtained using our combined noise-compensation and speaker-normalization method with the warp-factors obtained from clean-speech. Ideally, if noise compensation is perfect, then the two distributions should match. We define the following measure to quantify the deviation in the warp-factors for different approaches and for


c 0.8
HEQ 0.8
VTLN warped HEQ equalized feature
c 1.00
HEQ 1.0
c 1.2
HEQ 1.2
First pass Transcription
HEQ Model
Best VTLN wapred equalized features
VTLN+HEQ MODEL
T−VTLN Transformation matrix for 0.8 warp factor
c 1.00
HEQ 1.0
T 0.8
T 1.0
T 1.2
HEQ Equalized and T−VTLN Transformed feature
C 0.8 heq
C 1.0 heq
C 1.2 heq
First pass Transcription
HEQ Model
Best T−VTLN wapred equalized features
HEQ+T−VTLN MODEL

a) Train strategy for VTLN+HEQ

b) Train strategy for HEQ+TVTLN

Figure 2: Train strategies for VTLN+HEQ and HEQ+TVTLN

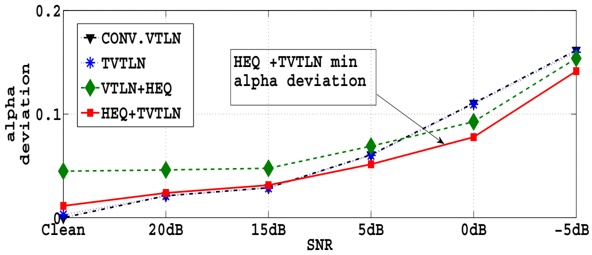


Figure 3: Alpha deviation for different approaches at various SNR levels

various SNR levels:

$$\alpha_{dev} = \frac{||\alpha_{clean} - \alpha_{expt}||_1}{length(\alpha_{clean})} \quad (3)$$

where α_{clean} and α_{expt} are vectors of warp-factors for all utterances obtained using clean conventional VTLN and the schemes considered in the previous section respectively. α_{dev} is a measure of deviation of the alpha values from the clean conventional VTLN warp-factors. Clean conventional VTLN warp-factors are considered to be the correct representation of speaker characteristics. The α_{dev} for conventional VTLN, T-VTLN, VTLN+HEQ and HEQ+TVTLN are plotted for degrading SNR's in Fig. 3. The plot for conventional VTLN overlaps with T-VTLN and hence is not clearly visible. α_{dev} for HEQ+TVTLN in clean conditions is 0.017 (less than one step, 0.02) indicating, that this approach results in warping that closely matches clean-speech warping. However, α_{dev} for VTLN+HEQ in clean conditions is 0.042, indicating changes in the speaker characteristics. Furthermore, the α_{dev} is large for conventional VTLN and T-VTLN, at low SNR's showing the impact of the noise on warp-factor estimation. Combining them with HEQ reduces α_{dev} and the minimum α_{dev} is obtained for the proposed HEQ+TVTLN approach.

4. Sub-band HEQ (S-HEQ)

We have proposed a novel modification to histogram equalization technique in our companion paper [4]. In S-HEQ, along with overall histogram, sub-band level histograms (low pass band and high pass band) are also equalized. S-HEQ is a two step equalization process. First, HEQ is applied on the conventional MFCC cepstra to get the HEQ equalized cepstra. This does not optimally equalize the sub-bands of cepstra. Hence a second level of equalization is done on the low-pass filtered (LPF) and high-pass filtered (HPF) components of HEQ equalized cepstra. LP and HP filtering is done by a simple averaging and differencing operation of adjacent cepstra within a each frame. Equalized LPF and HPF cepstra are then added to get a new set of features (with same dimension as MFCC) termed as S-HEQ features.

S-HEQ was shown to provide significantly better results over HEQ. Since there are two additional histogram equalizations, a combined approach of S-HEQ and T-VTLN seems most appropriate. This is implemented in similar fashion to HEQ+TVTLN, and gives significantly better results compared to other methods. Further, to illustrate its use in practice, SHEQ+TVTLN is implemented using *conventional* MFCC with no change in filter-bank unlike the previous experiments. Therefore, T-VTLN matrices operate on conventional 13-dimensional MFCC features to give 13-dimensional warped features. We again emphasize the advantage of T-VTLN for warping in the cepstra domain, which allows us to perform S-HEQ before VTLN.

5. Experimental Results

5.1. Experimental Set-up

The experiments are carried out on Aurora 2 and Aurora 4 database using HMM tool kit (HTK). The experiments had modifications in the signal processing steps as explained in Section 2. The modification was done to *match* the performance of T-VTLN with conventional VTLN. In particular, effort was made to match the warped cepstra and obtain similar warp-factor distribution. The front end has 40 filters, with the bandwidth equivalent to using 12 filters. SHEQ+TVTLN experiments were performed in *conventional MFCC framework*, where 23 filters were used. 13 dimension feature vectors with C0 instead of logarithmic energy are used for parametrization. Cepstral mean subtraction is performed by sentence-by-sentence mean subtraction of each cepstral coefficient. First and second order regression coefficients are augmented to get a 39 dimensional feature vector. For Aurora 2 database word level HMM model with 16 states with 6 Gaussians per state is used. For Aurora 4 database, cross word triphone HMM model is built with 3 tied states and 6 mixtures per state and standard WSJ bigram language model is used. Histogram Equalization is applied on the 13 cepstral coefficients (except for HEQ+TVTLN). In case of HEQ+TVTLN, 21 cepstra are equalized and then T-VTLN are applied to transform it to 13 dimensional equalized and warped features. Regression coefficients are calculated after equalization. During training, 13 reference CDF's are built, one for each feature element by averaging over the whole clean train utterances. Train and test utterances are equalized using

Test Case	T-01	T-02	T-03	T-04	T-05	T-06	T-07	T-08	T-09	T-10	T-11	T-12	T-13	T-14	Avg	R.I%
Baseline	88.59	76.27	52.79	54.55	46.95	56.87	46.31	77.53	64.79	44.70	42.76	35.89	47.28	36.75	55.15	0
HEQ	89.05	75.98	59.74	61.18	60.98	63.27	58.94	78.78	66.08	48.29	49.67	45.25	52.33	47.43	61.21	11.0
VTLN	90.42	82.29	58.81	59.54	54.14	63.48	51.60	81.32	71.62	51.45	48.31	42.46	54.49	43.96	60.99	10.60
T-VTLN	90.36	82.18	58.81	59.48	54.29	63.14	51.78	81.28	71.53	51.11	48.40	42.52	54.29	43.66	60.92	10.47
VTLN+HEQ	90.29	81.47	67.18	67.42	65.53	69.94	63.65	80.95	71.79	56.04	54.98	51.30	59.85	55.00	66.81	21.16
HEQ+TVTLN	90.57	81.97	66.84	66.58	64.86	69.21	63.55	81.66	72.22	55.69	54.47	50.55	59.20	54.44	66.56	20.69
S-HEQ (Conv MFCC)	88.51	81.56	68.22	64.08	65.08	69.64	64.69	82.42	76.69	62.10	57.89	54.94	63.27	58.88	68.43	24.1
S-HEQ+T-VTLN (Conv MFCC)	91.03	86.42	74.29	69.68	69.72	74.69	70.28	86.76	81.71	68.43	62.82	61.35	69.87	64.02	73.65	33.54

Table 1: Aurora-4 recognition accuracy results for different additive noise types (T-02 to T-07) and convolutive and additive noise type (T-08 to T-14). T-01 shows the recognition accuracy for clean speech

	BaseLine	HEQ	VTLN	T-VTLN	VTLN+HEQ	HEQ+TVTLN	S-HEQ (Conv MFCC)	S-HEQ+TVTLN (Conv MFCC)
clean	99.31	99.14	99.45	99.45	99.39	99.39	99.17	99.42
20 dB	97.92	97.80	98.15	98.12	98.24	98.41	97.76	98.50
15 dB	94.59	95.65	95.05	95.10	96.37	96.68	95.75	97.05
10 dB	83.06	89.85	83.74	83.71	90.78	91.56	90.83	93.22
5 dB	54.34	75.32	54.02	54.26	77.31	77.85	78.07	82.33
0 dB	25.13	44.28	21.91	22.07	44.54	45.40	51.79	55.11
-5 dB	12.91	15.16	10.47	10.56	14.86	16.03	21.66	21.73
Avg	71.01	80.58	70.57	70.65	81.45	81.98	82.84	85.24
R.I %	0	13.4	-0.07	-0.07	14.7	15.5	16.70	20.03

Table 2: Recognition Results on Aurora 2 database

reference CDF's.

5.2. Discussion

Tables 1 and 2 show the results for the Aurora-4 and Aurora-2 database. Combining VTLN and HEQ has shown relative improvement of 21.6% and 14.7% for Aurora-4 and Aurora-2 respectively. Both VTLN+HEQ and HEQ+TVTLN show similar recognition results. HEQ+TVTLN is computationally efficient and has the minimum α_{dev}. HEQ+TVTLN has shown improvements for all noise levels and at all SNR's over baseline.

5.3. S-HEQ and T-VTLN in Conventional Signal Processing framework

We now show the combination of S-HEQ+T-VTLN with no modification in filter-bank to illustrate the use of our proposed method in practice. S-HEQ+TVTLN experiments were performed *in conventional MFCC framework* without any modifications in the signal processing steps. Significant relative improvement of 20% and 33.54% in recognition accuracy over baseline for Aurora-2 and Aurora-4 database was obtained. VTLN+HEQ implemented in conventional MFCC framework showed a word error rate (WER) of 17.72% and 31.63% for Aurora-2 and Aurora-4 databases respectively [3]. S-HEQ+TVTLN outperforms VTLN+HEQ in conventional framework with relative improvement of 16.7% and 16.8% in WER over VTLN+HEQ for Aurora-2 and Aurora-4 databases respectively.

6. Conclusion

This paper describes an efficient approach to overcome speaker and noise variability. VTLN+HEQ technique performs better than VTLN or HEQ alone, however has disadvantages as discussed in Introduction. In comparison with VTLN+HEQ, HEQ+TVTLN needs only one set of equalization, making it an efficient implementation. T-VTLN being used in-place of conv. VTLN adds to computational advantage. Thus, HEQ+TVTLN is computationally efficient, easy to implement and has similar recognition performance as VTLN+HEQ. HEQ+TVTLN show a minimum deviation in warp-factor distribution. We then exploit the convenience of T-VTLN (applied in cepstral domain) and combine with S-HEQ (improved approach to HEQ). S-HEQ+TVTLN on top of computational advantage, has a very good recognition accuracy.

7. Acknowledgements

This work was supported under the Indo-Spanish Joint Program of Co-operation in Science and Technology. The Indian group is supported under project DST/INT/SPAIN/P-5 of Ministry of Science and Technology. The Spanish Group is supported under project ACI2009-0892 by the Ministry of Science and Innovation.

8. References

[1] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Process*, no. 6, pp. 49–60, 1998.

[2] A. de la Torre, A. Peinado, J. Segura, J. Perez-Cordoba, M. Benitez, and A. Rubio, "Histogram equalization of speech representation for robust speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 355 – 366, May 2005.

[3] L. Garcia, C. Benitez, J. C. Segura, and S. Umesh, "Combining speaker and noise feature normalization techniques for automatic speech recognition," in *Proc. of ICASSP-2011*, May 2011.

[4] V. Joshi, R. Bilgi, S. Umesh, L. Garcia, and C. Benitez, "Sub band level histogram equalization for robust speech recognition," Submitted to Interspeech 2011.

[5] J. McDonough, W. Byrne, and X. Luo., "Speaker normalization with all-pass transforms," in *Proc. ICSLP*, 1998.

[6] M. Pitz, "Investigations on linear transformations for speaker adaptation and normalization," Ph.D. dissertation, RWTH Aachen, Mar. 2005.

[7] S. Panchapagesan, "Frequency warping by linear transformation of standard mfcc," in *Proc. Interspeech*, 2006.

[8] S. Umesh, A. Zolnay, and H. Ney, "Implementing frequency-warping and vtln through linear transformation of conventional mfcc," in *Interspeech*, 2005, pp. 269–272.

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