COMBINING SPEAKER AND NOISE FEATURE NORMALIZATION TECHNIQUES FOR AUTOMATIC SPEECH RECOGNITION

L. García, C. Benítez, J.C. Segura *

Dpt. of Signal Theory, Telematics and Communications
University of Granada, Spain

S.Umesh †

Dpt. of Electrical Engineering
IIT-Madras, India

ABSTRACT

This work deals with strategies to jointly reduce the speaker and environment mismatches in Automatic Speech Recognition. The consequences of environmental mismatch in the performance of conventional Vocal Tract Length Normalization algorithm are analyzed, observing the sensitivity of the warping factor distributions to the SNR fall. A new combined speaker-noise normalization strategy which reduces the effect of noise in VTLN by applying Histogram Equalization is proposed and experimented in AURORA2 and AURORA4 databases. Solid results are obtained and discussed to analyze the effectiveness of the described technique.

Index Terms— Speaker Normalization - Noise Reduction - Combined Strategies - VTLN -HEQ

1. INTRODUCTION

Best results in an Automatic Speech Recognition are obtained when the testing is done under conditions that are identical to those in which the recognition system was trained: however this is an ideal situation and, in real applications, it will almost never happen. In a real scenario there are two main sources of variability which produce mismatches between the training and test conditions: the first one is the inter-speaker variability in the spectra for the same enunciated sound, which significantly degrades the performance when many different speakers use the system. The second one is the effect of the environmental noise on the speech representation. Noise introduces a distortion in the feature space and, due to its random nature, it also causes a loss of information. In the literature these two sources of degradation have been well explored separately.

The approaches that address the speaker variability problem in ASR fall into two broad groups known as speaker adaptation and speaker normalization techniques. In the speaker adaptation category, the parameters of the acoustic models of the independent speaker system are adapted to a new speaker using some of his data. The speaker normalization approach takes into account the fact that the variation in the vocal tract lengths is considered a major source of speaker variability, and the most common approach to overcome it is called vocal tract length normalization (VTLN) [1], [2],[3]. While speaker adaptation methods usually estimate matrices, VTLN estimates only one parameter and is therefore useful for situation counting on less adaptation data. In practice VTLN implementations consisting of linear transformations applied directly on the MFCC feature domain [4], [5], [6] are the most used due to their lower computational cost and comparable performance

On the other hand, the strategies used to reduce the environmental noise are known as Robust Recognition Techniques; these methods are mainly focused on the minimization of the mismatch caused by noise. Some of these approaches consist of the adaptation of the models to the new environment; other methods try to compensate the effect of the noise over the acoustic signal or its parameterization, providing an estimation of the clean speech representation. Other set of techniques investigate how to characterize the speech signal using features with a behavior less sensitive to noise. Within this latest category, feature normalization algorithms define linear and non linear transformations to modify the noisy features statistics and make them equal to those of a reference set of clean data. One fundamental representative of the feature normalization set of techniques is Histogram Equalization (HEQ) [7] [8], which provides a transformation mapping the histogram of each component of the feature vector onto a reference histogram. Such transformation eliminates the non-linear effect of noise distorting the original feature space.

Some analysis has been done on how the performance of the speaker normalization techniques degrades in the presence of noise. R. Rose and A. Keyvani [9],[10] point at such degradation of VTLN algorithm and propose the usage of class dependent warping distributions associated to individual HMMS states; A. Miguel [11] proposes an augmented state space acoustic decoding method for speech variability normalization and reports results on noisy databases. D.R. Sanand [12] shows results of a linear transformation variation of VTLN for AURORA4 database.

This work advances in the analysis of the warping factor sensitivity to the speech signal SNR. Based on such high sensitivity observed, the combination of speaker normalization and environment robustness strategies is proposed. A novel approach that consists of merging optimally the noise normalization technique Histogram Equalization and the Speaker normalization technique VTLN is suggested.

The work is organized as follows: section 2 analyzes the effect of noise in the speaker normalization process and supports it with empirical data. Section 3 describes the strategy proposed. Section 4 describes the experimental work and results obtained. Conclusions and hints about future work are exposed in section 5.

2. EFFECT OF NOISE IN THE VOCAL TRACT LENGTH NORMALIZATION PROCESS

All of the many existing implementations of VTLN aim to find the optimal warping factor α to warp the frequency axis of the speech signal so that the spectra of different speakers uttering the same sound appear similar. The conventional method to implement such

*This work was supported under the Indo-Spanish Joint Programme of Cop-operation in Science & Technology. The Spanish Group is supported under project ACI2009-0892 by the Ministry of Science and Innovation.

†This work was supported under the Indo-Spanish Joint Programme of Cop-operation in Science & Technology. The Indian group is supported under project DST/INT/SPAIN/P-5 of Ministry of Science & Technology.

speaker normalization is based on the production of the warped features for all the possible α values in a given range, followed by a maximum likelihood based search over them to estimate the optimal α value. The optimal warp-factor estimation $\hat{\alpha}_i$ for an utterance *i* is given by:

$$\hat{\alpha}_i = argmax_\alpha Pr(C_i^\alpha|\lambda, W_i) \quad (1)$$

Where C_i^α represents the warped features of the i^{th} utterance and λ stands for the speaker independent acoustic model. W_i corresponds to the utterance's true transcription for training utterances. For test utterances, since the test true transcriptions are unknown, W_i is obtained as output of the first-pass recognition process. The likelihood is then calculated for all values of α from which the optimal $\hat{\alpha}$ will be chosen. Independent of which VTLN implementation is used, the performance of the speaker normalization is known to degrade for tasks where ambient acoustic noise is a significant source of variability. Many authors, [11], [10] have pointed at the need to overcome the strong influence of noise in the optimal warping factor estimation.

Figure 1 shows AURORA2 [13] accuracy results comparing the baseline parameterization scheme to the VTLN speaker normalization scheme in different noise conditions. AURORA2 gives very good results for clean baseline recognition and therefore leaves VTLN with very little margin to produce benefits. As the SNR decreases, the graph shows the tendency of VTLN to have worse performance than the baseline, and therefore the need to deal with noise when applying speaker normalization techniques. Figure 2 shows the optimal warping factor values distribution when applying VTLN for AURORA2 at different noise levels. Sub-figure *a)* presents the distribution for clean test utterances showing the two expected modes at both sides of the unit warping factor corresponding to the two genders. As SNR decreases (sub-figure *b)*) the bi-modal distribution is lost, to end grouping the warping factors towards the lowest values (see sub-figures *c)* and d)). Such behavior in low SNR is due to the increase of energy in the high-frequency band when compared to clean-speech. Since there is higher energy at high frequencies, VTLN will try to somehow *compress* it, that is, an aggressive compressing warping factor like 0.8 will be chosen.

The same warping factor distribution analysis has been done for AURORA4 database [14] and is shown in figure 3. Warping factor distributions have been shown for clean test (sub-figure *a)*), car noise as an example of stationary additive noise (sub-figure *b)*), non-stationary street noise (sub-figure *c)*), plus these three just mentioned tests adding them convolutive channel noise in the subsequent sub-figures *d)*, *e)* and *f)*. In this case, the expected clean warping factor distribution is mainly maintained for all test cases. The reason for this behavior is that the average AURORA4 SNRs per noisy experiment is 15 dB that, in agreement with AURORA2 behavior, is a SNR level for which VTLN produces *healthy* optimal warping factors.

3. COMBINED STRATEGY: PERFORMING VTLN IN A NOISE-ROBUST DOMAIN

The strategy proposed in this work consists of combining the speaker normalization and the noise normalization techniques in order to make the first one more robust against environmental mismatches. Histogram Equalization [8], [15] will be applied together with a basic VTLN implementation. HEQ feature normalization provides a non-linear transformation that can be seen as an extension of the cepstral mean and variance normalization, able to compensate the non-linear effects of noise not eliminated by those. Its proved efficiency added to its versatility to compensate different types of noises

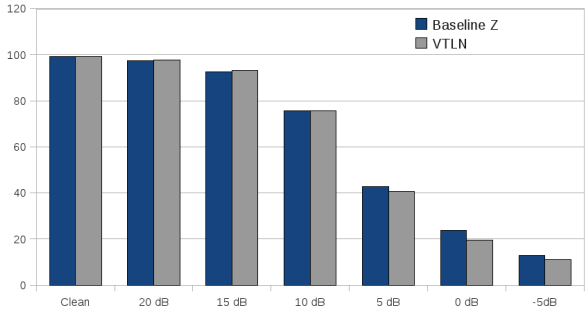


Fig. 1. *Baseline and VTLN Accuracy results for AURORA2 Database for different noise levels.*

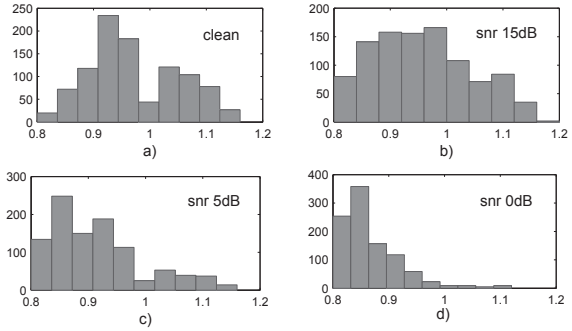


Fig. 2. *Warp factor distribution for different SNR in AURORA2: clean (sub-figure a), 15 dB (sub-figure b), 5 dB (sub-figure c) and 0 dB (sub-figure d).*

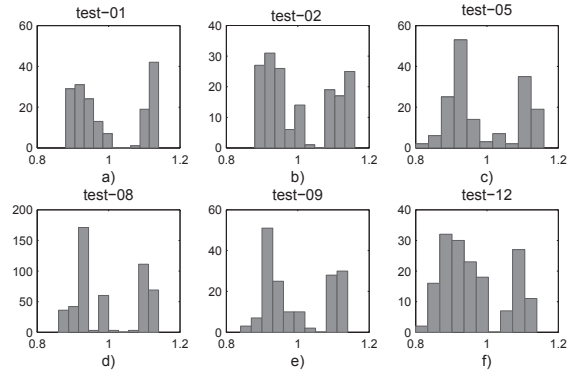
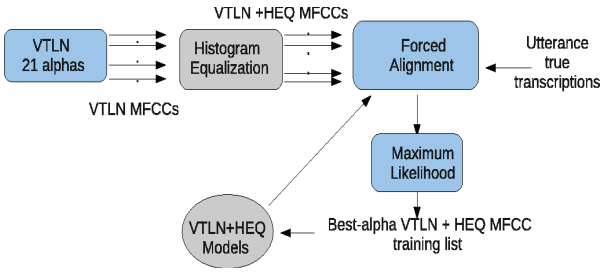


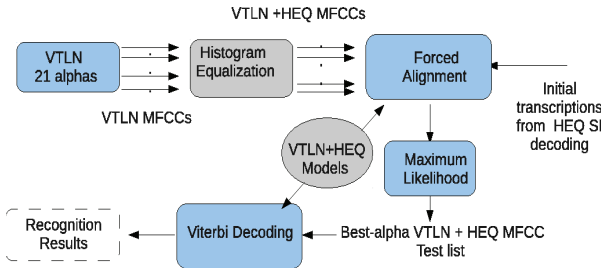
Fig. 3. *Warp factor distribution for different tests in Aurora 4 in the different sub-figures: a) clean speech, b) stationary additive noise, c) additive non-stationary noise, d) convolutive noise, e) conv.noise plus additive stationary noise, f)conv. noise plus additive non-stationary noise*

without any prior knowledge nor model, make it a very suitable approximation to be introduced in the joint scheme under exploration.

Figures 4(a) and 4(b) present a block diagram of the combined strategy proposed to merge conventional VTLN and HEQ normalization. The objective of the proposed structure is to eliminate the influence of noise in the process of warping factor selection. An initial preliminary step of the method proposed is to generate an speaker



(a) *Train strategy for the combined algorithm*



(b) *Test strategy for the combined algorithm.*

independent recognition system using HEQ normalized features (SI HEQ MODEL) that will be utilized during the train and test phases. Figure 4(a) presents the block diagram of the training phase in which the original VTLN features are equalized before using a VTLN-HEQ generated model to perform a forced alignment. Using a Maximum Likelihood criterion, the best warping factor are chosen from the forced alignment results forming a VTLN-HEQ training list that is used to re-estimate optimized VTLN-HEQ Models. This training procedure is iterated 3 times, using the speaker independent HEQ models (SI HEQ MODELS) as initial VTLN-HEQ Models in iteration 1.

Figure 4(b) describes the proposed recognition process which can be split into two decoding steps. A first decoding in done using SI-HEQ models to generate transcriptions that will be used to perform a forced alignment using the VTLN-HEQ models to select the best warping factor per utterance. Such best-alpha codified utterances are decoded in a second step using the VTLN-HEQ model to produce the final recognition results.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

The proposed scheme has been tested on AURORA2 [13] and AURORA4 [14] databases, following the standard clean training experiments. All the procedures for recognition and training are identical to the reference experiments with the exception of the front-end that includes the noise and speaker normalization procedure and HMMS re-estimations described in this research. The recognition system used for AURORA4 is based on continuous cross-word triphone models with 3 tied states and a mixture of 6 Gaussians per state. Training and recognition have been performed using the HMMTool Kit (HTK) software. The language model is the standard bi-gram for the WSJ0 task. In the case of AURORA2 connected digits task, each digit is modelled as a left to right continuous density HMM with 16 states and 6 Gaussians per state. A feature vector of 13 cepstral coefficients is used as the basic parameterization of the speech signal using C0 instead of the logarithmic energy. This basic feature vector is augmented with first and second order regressions yielding a final 39 components feature vector. The baseline reference system uses sentence-by-sentence subtraction of the mean values of each cepstral coefficient. For the proposed combined normalization technique (HEQ*VTLN), the parameters of the HEQ reference distribution have been obtained by averaging over the whole clean training set of utterances. Both training and test utterances have been then equalized to this reference distribution. Cepstral coefficients are equalized before the computation of the regressions.

4.2. Discussion

Table 2 presents the numerical accuracy results of the studied algorithms for AURORA2 database in terms of noise levels. The general behavior is a drop in the accuracy when the SNR is reduced. HEQ overcomes the noise very efficiently providing a baseline relative improvement of 22.85%. In clean conditions, the very good results of AURORA2 baseline reduce VTLN's margin to improve. VTLN slightly ameliorates baseline up to SNRs of 15 dB. For lower SNRs the noise degrades VTLN performance for the reasons analyzed in section 2. The joint method described in section 3 produces the best accuracy results for all the noise levels giving a baseline relative improvement of 23.93%. This behavior shows the mutual benefits that each individual normalization technique provides the other with.

Table 4.2 shows the experiments performed using AURORA4 database. The average SNR for each of AURORA4 tests is 15 dB. For these reason all the methods analyzed improve the baseline parameterization. The combined strategy HEQ*VTLN obtains the best results for all the test (23.90% baseline relative improvement), pointing again at the mutual positive interaction between speaker and noise normalization techniques. Clean scenario results underline the robustness of VTLN approach. HEQ also improves baseline in clean conditions because its philosophy of equalizing to a reference distribution calculated with the whole clean training data set implies a kind of *blind* speaker normalization. Applying the combined HEQ*VTLN still improves both separate techniques as the warping factor estimation becomes more reliable. Test 2 (stationary additive car noise) and Test 9 (stationary additive car noise plus channel convolutive noise) deserve a special mention as VTLN outperforms HEQ when applied separately. This fact could mean that for certain normalization algorithms the boundaries between speaker and noise mismatches might not so rigid.

	Baseline	**VTLN**	**HEQ**	**HEQ*VTLN**
Clean	99.11	99.30	99.03	99.18
20 dB	97.29	97.80	97.61	98.12
15 dB	92.55	93.24	95.62	96.43
10 dB	75.60	75.68	90.38	91.44
5 dB	42.82	40.56	76.92	77.88
0 dB	23.69	19.47	47.26	47.52
-5 dB	12.92	10.98	17.34	17.64
Average	66.39	65.35	81.56	82.28
Rel. Improv.	**0%**	**-1.57%**	**22.85%**	**23.93%**

Table 2. AURORA2 accuracy results for different noise levels.

	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	Avrg.	R-I(%)
Baseline	87.15	74.33	55.87	54.03	47.81	56.43	47.03	75.84	64.53	45.45	42.39	35.87	47.55	38.23	55.18	0
VTLN	89.24	82.36	61.99	59.56	53.73	61.44	51.09	81.99	71.38	51.31	47.07	39.63	54.25	43.86	60.64	9.89
HEQ	88.18	77.27	62.21	61.10	63.06	62.62	59.01	78.71	67.37	53.92	50.64	49.39	55.51	50.24	62.81	13.83
H-VTLN	90.68	83.13	68.47	68.69	68.29	69.61	64.31	81.73	73.96	59.30	57.62	53.33	61.80	56.28	68.37	23.90

Table 1. AURORA4 accuracy results for clean speech (Test 01), different additive noise types (Test 02 to Test 07) and convolutive and additive noise (Test 08 to Test 14). H-VTLN stands for the combined HEQ VTLN normalization.

5. CONCLUSIONS AND FUTURE WORK

This work identifies the convenience of merging complementary mismatch reduction techniques like speaker and noise normalization. An analysis of the effect of noise in the conventional VTLN implementation has been done focusing on the noise effects over the warping factor estimation under low noise conditions: α tend to move to lower values because VTLN interprets noise as high frequency acoustic information. A combined algorithm has been proposed and experimented producing good recognition results that confirm the mutual benefit of two families of techniques. The extension of this analysis to VTLN implementations based on linear transformations with lower computational cost than conventional VTLN is a very convenient next step.

6. REFERENCES

[1] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[2] P.Zhan and A. Waible, "Vocal tract length normalization for large vocabulary continouous speech recognition," *Technical report CMU-CS-97-148*, 1997.

[3] L. Lee and R. Rose, "Frequency warping approach to speaker normalization," *IEEE Transactions on Speech and Audio Processing*, 1998.

[4] S. Umesh, Andras Zolnay, and Hermann Ney, "Implementing frequency-warping and vtln through linear transformation of conventional mfcc," *Proc. INTERSPEECH-2005*, 2005.

[5] D. R. Sanand, D. D. Kumar, and S. Umesh, "Linear transformation approach to vtln using dynamic frequency warping," in *Proc. of INTERSPEECH-2007*.

[6] 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 sufficient statistics," in *Proc of INTERSPEECH-2008*.

[7] A. De la Torre, J.C. Segura, C. Benítez, A. Peinado, and A. Rubio, "Non-linear transformation of the feature space for robust speech recognition," in *Proc. of ICASSP'02*.

[8] A. de la Torre, A. Peinado, J. C. Segura, J.L. Pérez Córdoba, C. Benítez, and A. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, 2005.

[9] A. Keyvani, "Robustness in asr: An experimental study of the inter-relationship between discriminant feature-space transformation, speaker normalization and environment compensation," *Ph.D. dissertation, McGill University*, 2007.

[10] R.C. Rose, A. Miguel, and A. Keyvani, "Improving robustness in frequency warping based speaker normalization," *Signal Processing Letters, IEEE*, vol. 15, pp. 225–228, 2008.

[11] A. Miguel, E. Lleida, R. Rose, L. Buera, O.Saz, and A. Ortega, "Capturing local variability for speaker normalization in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, n 3, pp. 578–593, March 2008.

[12] D.R. Sanand, R. Schlueter, and H. Ney, "Revisiting vtln using linear transformation on conventional mfcc," *Interspeech'2010*, 2010.

[13] D. Pearce and H.G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ICSLP 2000*.

[14] H.G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task," Tech. Rep., STQ AURORA DSR Working Group, 2002.

[15] F. Hilger and H.Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Transactions on speech and audio processing*, 2006.