ROBUST SPEECH RECOGNITION THROUGH SELECTION OF SPEAKER AND ENVIRONMENT TRANSFORMS

Raghavendra Bilgi, Vikas Joshi, S. Umesh

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

L. Garcia, C. Benitez

Dept of Signal Theory,Telematics & Communications
University of Granada, Spain
luzgm, carmen@ugr.es

ABSTRACT

In this paper, we address the problem of robustness to *both* noise and speaker-variability in automatic speech recognition (ASR). We propose the use of *pre-computed* Noise and Speaker transforms, and an optimal combination of these two transforms are *chosen* during test using maximum-likelihood (ML) criterion. These *pre-computed* transforms are obtained during training by using data obtained from different noise conditions that are usually encountered for that particular ASR task. The environment transforms are obtained during training using constrained-MLLR (CMLLR) framework, while for speaker-transforms we use the analytically determined linear-VTLN matrices. Even though the exact noise environment may not be encountered during test, the ML-based choice of the closest Environment transform provides "sufficient" cleaning and this is corroborated by experimental results with performance comparable to histogram equalization or Vector Taylor Series approaches on Aurora-2 task. The proposed method is simple since it involves only the *choice* of pre-computed environment and speaker transforms and therefore, can be applied with very little test data unlike many other speaker and noise-compensation methods.

Index Terms— speaker adaptation, environment adaptation, robustness

1. INTRODUCTION

Speaker-variability and Noise are two major sources of performance degradation in ASR. In many practical ASR applications, both of these sources of degradation are present, and hence there is a lot of interest in reducing both. However, historically, most of the research work have often focused on reducing only one of these sources without taking into account the other problem. For example, most work on reducing inter-speaker variability often focus only on this problem, without taking into account the problem of noise. Two broad approaches to speaker-normalization are speaker-adaptation based approaches such as Maximum Likelihood Linear Regression (MLLR) or Constrained-MLLR (CMLLR) [1] and Vocal-tract Length Normalization (VTLN) [2]. Usually, these approaches are "two-pass" in nature, where the first-pass recognition output is used to estimate the transform parameters or VTLN-warping factor, before the final recognition is done. Since MLLR/CMLLR require estimation of the parameters of the transformation matrix, they require larger amount of adaptation data (about 30 sec.) for robust estimation of the parameters when compared to VTLN which requires only the warp-factor estimation. The parameter estimates are sensitive to first-pass transcription errors (more so in speaker-adaptation cases than VTLN), and hence the presence of noise could affect the efficacy of these speaker-normalization methods.

Similarly, noise-compensation algorithms have been proposed, which often do not account for inter-speaker variability. Two commonly used noise-compensation approaches are those based on histogram equalization (HEQ) [3] and those based on Vector Taylor Series (VTS) [4]. In the histogram based approaches, adequate speech data is required to get robust estimates of the quantiles, while in the VTS based approach the noise models are obtained from the first few and last few frames of the utterance.

Recently, there has been lot of interest in reducing both the noise and speaker-variability. Various combinations of speaker and noise adaption techniques have been studied and are shown to mitigate the effects of noise and at the same time compensate for speaker variability. Combination of VTLN with HEQ is studied in [5][6]. Combination of VTS with VTLN [7] and VTS with MLLR are studied in [8]. Gales [9] proposed the acoustic-factorization approach to separate the noise and speaker effects and uses cluster-adaptive (CAT) approach for environment transform estimation and MLLR for speaker-transform estimation. In [10] cascade of CMLLR transforms are used which enables the use of transform estimated in one environment to be used with same speaker in another environment. However, the paper assumes the knowledge of speaker and noise environment. Increasingly speech recognition systems are used in mobile devices where environment changes quite often and the SNR levels of the speech signal also changes significantly. In such a case, a single transform to compensate the noise variability for that environment may not be appropriate. Further, in [9][10], the noise and environments transforms have to be *estimated* using test utterances as adaptation data.

In this paper, we propose the estimation of noise transforms *during the training step*. These noise transforms are essentially CMLLR transforms that are applied on the features and represent a "cleaning" of the features. The basic premise is that the different noise conditions that are encountered in a particular ASR task are captured by these noise-transforms and stored. Therefore, these matrices are "pre-computed" and one of these are chosen during the test step using a maximum-likelihood criterion. Even if a noise condition that is not "seen" during training appears, it is hoped that a transform that closely matches the condition will be chosen from among the pre-computed transforms enabling the cleaning of the features and providing improvement in the recognition accuracy. Similarly, to reduce inter-speaker variability, we use a set of pre-computed Linear-VTLN matrices corresponding to the range of warp-factors used in the ML search. In our case, we use 21 warp-factors, and hence the 21 Linear-VTLN warp-matrices are computed using our analytically determined Linear-VTLN approach [11]. The advantage of the use of such pre-computed noise and speaker transforms is that during

Fig. 1. Estimation of Environment and Speaker Transform

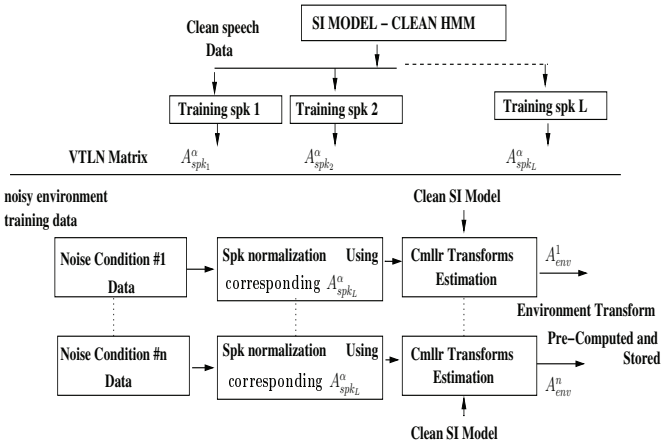
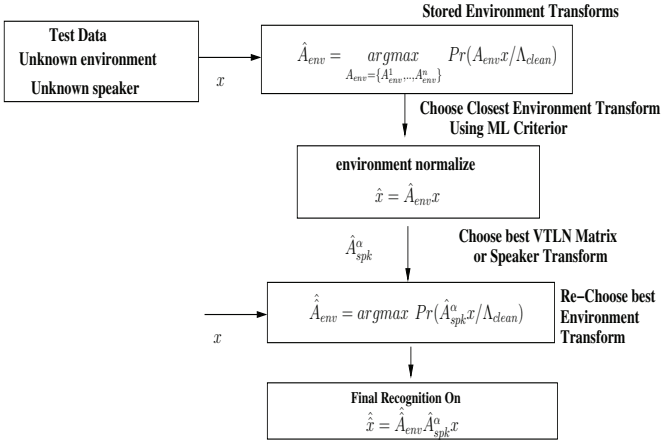


Fig. 2. Adaptation during test using pre-computed transform



test, we need to choose only one of these matrices, and hence very little "adaptation" data is required to make the robust choice. While in this paper we discuss using these transforms for adaptation of test data, this can be easily extended to adaptive training.

The rest of the paper is organized as follows. In Section 2, we discuss the approach to compensate the noise and speaker variability. Experiments to evaluate the performance of the proposed approach is discussed in Sections 3.1 and 3.2. Finally, we conclude the paper in Section 4.

2. PROPOSED APPROACH USING ENVIRONMENT AND SPEAKER TRANSFORM

2.1. Estimation of Environment and Speaker Transform

Fig. (1) shows the block diagram of the proposed environment transform estimation process. Using clean Speaker Independent (SI) HMM model as the baseline, speaker specific transforms are first estimated from relatively *clean* speech utterance from among all the utterances in the same environment. By choosing clean utterances, the transform estimated would be a good representation of speaker characteristics. These speaker specific transforms are then used to normalize the features so as to remove the speaker related variability as show in Eq. (1)

$$y^i_{train} = A^{\alpha}_{spk} x^i_{train} \tag{1}$$

Where α is the VTLN warp factor and A^{α}_{spk} is linear-VTLN speaker transform matrix.

Once the speaker variability is removed from the features we can use all the train utterances collected in a specific noise environment (e.g. car noise, restaurant etc.) at different noise levels (e.g very noisy, noisy, less noisy, clean) and estimate environment noise specific CMLLR transforms. The corresponding auxiliary function is shown in Eq. (2):

$$Q(A_{env}, \bar{A}_{env}) = \sum \gamma^i_m \; log \; Pr(A_{env} y^i_{train} / \Lambda_{clean}) \tag{2}$$

where the summation is over all utterances, frames and Gaussian components associated with that noise environment and Λ_{clean} is clean SI HMM model. These transforms will capture only the noise variability as they are estimated from speaker normalized features.

2.2. Adaptation During Test Using Pre-Computed Transforms

Fig. (2) shows the block diagram of the proposed transform selection and recognition process. Selection of the Environment transform is done in ML framework as shown in Eq. (3). Where A^n_{env} is the nth pre-computed environment specific CMLLR transforms obtained during training. Selection of transforms can be done more efficiently by using sufficient statistics as in [12]. However in our implementation we have computed individual likelihood for each transform.

$$\hat{A}_{env} = \underset{A_{env}=\{A^1_{env},\ldots,A^n_{env}\}}{argmax} Pr(A_{env} x / \Lambda_{clean}) \tag{3}$$

These environment normalized features using $\hat{A}_{env}$ are then used to estimate the speaker specific transform. We *first* do the environment normalization as VTLN warp factor estimation is sensitive to noise [5][6]. The noise *cleaned* features are then used to estimate the speaker transform. Eq. (4) shows the selection of VTLN transform obtained after the application of environment transform, i.e.

$$\hat{\alpha}_{ML} = \underset{\alpha}{argmax} \; Pr(A^{\alpha}_{spk} \hat{A}_{env} x_i / \Lambda_{clean}, T_i), \tag{4}$$

where T_i is the first-pass transcription.

The Speaker normalized features are then used to *re-choose* the environment transforms as shown in Eq. (5):

$$\hat{\hat{A}}_{env} = \underset{A_{env}=\{A^1_{env},\ldots,A^n_{env}\}}{argmax} Pr(A_{env} \hat{A}^{\alpha}_{spk} x / \Lambda_{clean}), \tag{5}$$

where $\hat{A}^{\alpha}_{spk}$ represents optimal speaker transform obtained from Eq. (4). Environment and speaker normalized features obtained from Eq. (6) are used for recognition.

$$\hat{x}^i_{test} = \hat{\hat{A}}_{env} \hat{A}^{\alpha}_{spk} x^i_{test} \tag{6}$$

Fig. 3. Comparison of recognition accuracy for set A between (i). single transform per environment (ii). True SNR specific transform and (iii). SNR transforms selected in ML framework

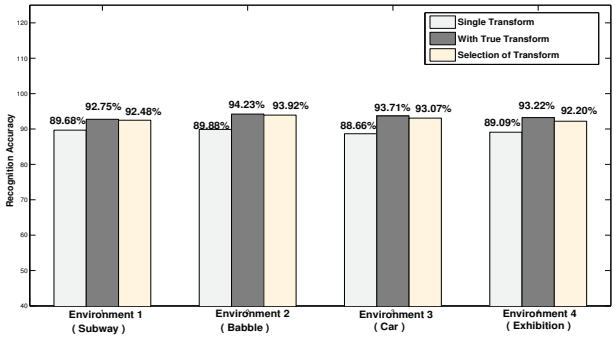
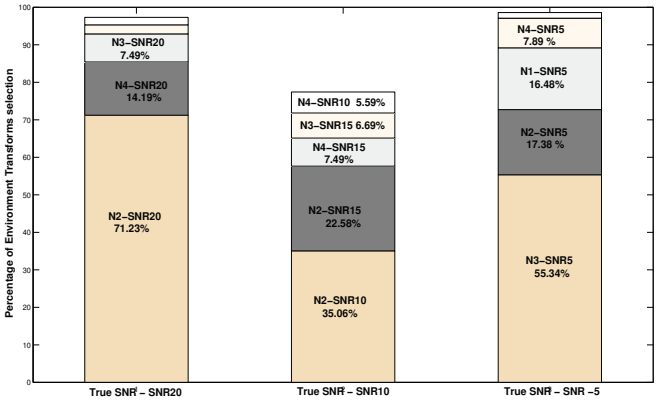


Fig. 4. Shows the preference of transforms selected in ML framework for N1 environment in test set A for 1001 utterances per SNR. Where N2, N3, N4 specify the other environment in Set A for which environment transforms are estimated



3. EXPERIMENTAL RESULTS

3.1. Experimental Setup

In our experiment, Aurora 2 database is used which consists of speech utterances with eight types of environments. Training set has speech utterance for four different environments at SNRs between 20dB and 5dB. Evaluation is performed using two test sets each of which has SNRs between 20 dB and -5 dB. Set A contains noise types seen in the training data while Set B has noise types *unseen in the training data*.

Acoustic models were trained from clean training data using HTK. An HMM with 18 states per digit and 3 mixtures per state is created for each digit as a word model. There is a three state silence model with 6 Gaussian per state and one state short pause model tied to the middle state of the silence. Standard 39 dimension MFCC features consisting of 13 static, delta and delta-delta features were computed from power spectral observation and $C0$ was used instead of log energy. The baseline clean HMM system with cepstral mean normalization (CMN) had a word accuracy of 67.45% on Set A and 72.26% accuracy on Set B. Average accuracy is calculated without accounting the accuracy of *clean* and *SNR* -5.

3.2. Experiments and Results

The following experiments were performed to check the advantage of choosing the environment transform and to study the combination of noise and speaker transform. In the case of Aurora-2, we have different environments (e.g car noise, street noise, babble) as well as different noise levels i.e SNR levels. In the first set of the experiment, SNR specific environment CMLLR transforms were estimated for all the four training environments. Training set has utterances for four different environment and each environment has utterance collected in 5 different SNRs between $20dB$ and $5dB$. Using the data total of 20 SNR specific environment CMLLR transforms were estimated. Evaluation is done on set A and set B. Note that set B has noise conditions *not seen* during training. One transform among the 20 SNR specific transform is chosen as best environment and the recognition is done by transforming the feature using the best environment as in Eq. (3). To find the upper bound of the results, same experiment is repeated now with *true SNR specific* transforms applied on the test features. We also considered third case where one single transform is estimated per environment using all the SNR levels for that environment. Fig. (3) compares the result of transform selection approach on Set A with the ideal (true) case and with the single transform case. Test set has two additional SNR levels $0dB$ and $-5dB$. Since transform for these SNR's are not learned during training, we show average results over clean to SNR 5. From Fig. (3) we see that, compared to using single transform for one environment (e.g car noise etc) at all SNR levels (e.g clean, noisy, very noisy) using SNR specific transform provides significant improvement. Further, our proposed method of choosing transform without any knowledge of noise has *comparable* performance to using true SNR and environment specific transform.

Fig. (4) shows the true SNR (i.e ground truth) of the utterance and SNR transform actually selected in ML framework as in Eq. (3). From the figure it is clear that best transform were chosen from same or nearby SNR transforms. For example, when the SNR is $10dB$ selected transforms are from SNR $10dB$, SNR $15dB$. Similarly when the true SNR is $-5dB$, since the available SNR transforms are only from clean to SNR 5, it chooses the "closest" level namely SNR $5dB$. The results support our argument that ML approach chooses the appropriate transform which matches closely to the noise level.

Previous experiments did not account for the speaker variability. We now consider the selection of both environment and speaker transform. During testing, first select the best SNR specific environment transform and and then use it to clean the feature. Once the features are noise normalized, VTLN warp factor is estimated for the speaker w.r.t clean HMM model. The best Speaker transforms ($\hat{A}^{\alpha}_{spk}$) are later used to remove the speaker related variability from the feature and Environment transform is *re-chosen* on the speaker normalized features. Recognition is done on speaker normalized and environment normalized features as shown in Eq. (6).

3.3. Discussion of Recognition Results Using Environment and Speaker Transforms

Table. (1)(a) shows the recognition results of baseline clean HMM system for Set B. Note that noise environment for Set B is not seen during training and represents unseen environment. Results for the proposed approach with VTLN as speaker transform is shown in Table. (1)(b). Combination of environment transform with VTLN shows an impressive relative improvement of 17.3% over baseline. This result is comparable to the result achieved with combination of

Table 1. Recognition results on Set B for the proposed method with VTLN and CMLLR as speaker transforms

(a) Baseline CMN System						(b) Proposed : $y^i_{test} = \hat{\hat{A}}_{env}\hat{A}^{\alpha}_{spk}x^i_{test}$					(c) Upper Bound : $y^i_{test} = \hat{\hat{A}}_{env}\hat{A}_{spk}x^i_{test}$				
	Env 1	Env 2	Env 3	Env 4	Avg	Env 1	Env 2	Env 3	Env 4	Avg	Env 1	Env 2	Env 3	Env 4	Avg
Clean	99.23	99.24	99.08	99.44	99.25	99.17	99.06	98.96	99.29	99.12	99.26	99.30	99.14	99.44	99.29
20 dB	97.94	97.19	97.91	98.18	97.81	98.65	97.58	98.81	98.77	98.45	99.05	98.43	99.22	99.54	99.06
15 dB	95.03	93.14	95.29	94.85	94.58	97.36	96.40	97.49	97.28	97.13	98.59	98.00	98.72	98.89	98.55
10 dB	85.60	80.47	87.80	82.84	84.18	93.89	90.75	95.08	94.01	93.43	96.50	94.47	97.08	96.51	96.14
5 dB	62.08	51.51	62.24	52.79	57.16	82.19	74.82	84.58	80.78	82.09	88.92	85.10	90.28	87.41	87.93
0 dB	30.61	23.49	31.91	24.28	27.57	55.76	44.74	59.47	50.85	52.71	67.39	59.55	68.80	61.15	64.22
-5 dB	13.97	11.46	15.48	11.48	13.10	24.16	17.35	26.96	20.21	22.17	34.33	26.81	35.31	30.15	31.65
Avg	74.25	69.16	75.03	70.59	72.26	86.77	80.86	87.09	84.33	84.76	99.09	87.11	90.82	88.7	89.18

nonlinear compensation techniques like HEQ and VTS with VTLN [5][7]. Table. (1)(c) also shows the results for the best case (which is upper bound) where in CMLLR transform is used instead of VTLN as speaker transform. In this case, using noise cleaned data, CMLLR transform is estimated for each speaker, for every environment and every noise level. Therefore we do a separate noise adaptation at every noise level for every noise environment of each speaker. This represents upper bound and is shown in Table. (1)(c). Note that in environment selection in Table. (1)(b) there are no transforms available at SNR 0 and SNR -5 due to lack of training data. However for CMLLR as speaker transform, some amount of adaptation data is used for these SNR's and hence it shows significant improvement. At all other level, our proposed method is quite close to upper bound and significantly better than the baseline.

4. CONCLUSION

In this paper, we have shown that by selecting appropriate Environment and Speaker Transforms from a pre-computed set, we can achieve performance comparable to existing method such as Histogram Equalization. More importantly the method works equally well on test data with noise environment NOT seen during training. The pre-computed environment transforms were obtained from training data using CMLLR framework while the speaker transform are a set of Linear-VTLN matrices corresponding to the range of warp factors. Since these transforms need to be only "selected", they can be applied even when very little test data is available making it attractive when compared to other noise and speaker-adaptation methods.

5. ACKNOWLEDGEMENT

This work was supported in part under the SERC project funding SR/S3/EECE/058/2008 and DST/INT/SPAIN/P5 of Department of Science and Technology, India. The Spanish Group is supported under project ACI2009-0892 by the Ministry of Science and Innovation, Spain.

6. REFERENCES

[1] M.J.F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

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

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

[4] Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. ICASSP-96*, 1996, pp. 733–736.

[5] V. Joshi, R. Bilgi, S. Umesh, L. Garcia, and C. Benitez, "Efficient speaker and noise normalization for robust speech recognition," in *Proc. Interspeech*, 2011.

[6] 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*, 2011.

[7] K. K. Chin, Haitian Xu, Mark J. F. Gales, Catherine Breslin, and Kate Knill, "Rapid joint speaker and noise compensation for robust speech recognition," in *ICASSP*, 2011, pp. 5500–5503.

[8] Y. Q. Wang and M.J.F. Gales, "Speaker and noise factorisation on the aurora4 task," in *ICASSP*, 2011, pp. 4584–4587.

[9] M.J.F. Gales, "Acoustic factorisation," in *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, 2001, pp. 77 – 80.

[10] Michael L. Seltzer and Alex Acero, "Separating speaker and environmental variability using factored transforms," in *Proc. Interspeech*, 2011.

[11] Doddipatla Rama Sanand, *Linear Transform Approaches To Vocal Tract Length Normalization for Automatic Speech Recognition*, Ph.D. thesis, Indian Institute of Techonology, Kanpur, July 2009.

[12] P. T. Akhil, S. P. Rath, S. Umesh, and D. Rama Sanand, "A computationally efficient approach to warp factor estimation in vtln using em algorithm and sufficient statistics," in *INTERSPEECH*, 2008, pp. 1713–1716.