

Evaluation of the Combined Use of MEMLIN and MLLR on the Non-native Adaptation Task of Hiwire Project Database

Luis Buera, Antonio Miguel, Óscar Saz, Eduardo Lleida, Alfonso Ortega

Communication Technologies Group (GTC), I3A, University of Zaragoza, Spain.

{lbuera, amiguel, oskarsaz, lleida, ortega}@unizar.es

Abstract

This paper describes the performance of the combination of Multi-Environment Model-based Linear Normalization, MEMLIN, which provides an estimation of the uncorrupted feature vector, with Maximum Likelihood Linear Regression, MLLR, for the collected database under the auspices of the IST-EU STREP project HIWIRE. In this work the results for the non-native adaptation task (NNA) are presented. The HIWIRE project database consist on command and control aeronautics application utterances pronounced by non-native speakers which are digitally corrupted with airplane cockpit noise. Thus, three noise conditions are defined: low, medium and high noise. In the proposed system, each MEMLIN-normalized feature vector is decoded using the MLLR-adapted acoustic models. The experiments show that an important improvement is reached combining MEMLIN and MLLR methods for all kinds of non-native speakers and noise conditions.

Index Terms: Hiwire project, robust speech recognition, non-native adaptation task.

1. Introduction

When training and testing acoustic conditions differ, the accuracy of speech recognition systems rapidly degrades. To compensate this mismatch, classic robustness techniques have been developed along the following two main lines of research: acoustic model adaptation methods, and feature vector normalization methods. In general, acoustic model adaptation methods produce better results [1] because they can model the uncertainty caused by the noise statistics. However, these methods usually require more data and computing time than feature vector normalization methods do, which do not produce as good results but provide more on-line solutions. Hybrid techniques, which are the combination of a feature vector normalization method and an acoustic model adaptation method, also exist [2].

A previous work [3] shows that Multi-Environment Model-based Linear Normalization, MEMLIN, with cross-probability model based on GMMs is effective to compensate the effects of dynamic and adverse car conditions. MEMLIN is an empirical feature vector normalization based on stereo data and the MMSE estimator, with joint modelling of clean and noisy spaces by Gaussian Mixture Models (GMMs). Therefore, a bias vector transformation is associated with each pair of Gaussians from the clean and the noisy spaces. A critical point in MEMLIN is the estimation of the probability of the clean model Gaussian, given the noisy model one and the noisy feature vec-

tor (cross-probability model) which, in this work, is modelled with GMMs.

On the other hand, classic acoustic model adaptation methods, e.g. Maximum A Posteriori, MAP, [4], or Maximum Likelihood Linear Regression, MLLR, [5] take into account all kinds of degradations of the feature vectors by mapping the parameters of the acoustic models to the noisy space.

The HIWIRE project database was designed to be a tool for the testing of robust speech recognition techniques. In this database, the mismatch between clean and noisy conditions is modelled with the artificial addition to clean signals of airplane cockpit noise in three Signal-to-Noise Ratios (SNRs), which defines three noise situations: low, medium and high. The non-native adaptation task (NNA) of this database is a good set to measure the performance of the combination of feature vector normalization methods with acoustic model adaptation methods. This subset of the database allows to work with model adaptation methods as it uses half of the speaker utterances for training while the other half is used for testing; furthermore, the NNA task contains stereo data on the train set to allow the training of empirical feature vector normalization methods.

In this work we propose an hybrid solution for the non-native adaptation task which combines MEMLIN with cross-probability model based on GMMs with MLLR. Thus, normalized space for each speaker and noise condition is modelled with MLLR, so that in recognition, each MEMLIN-normalized feature vector is decoded with the corresponding adapted acoustic models, which are generated with the noisy adaptation feature vectors which have been normalized previously.

This paper is organized as follows: In Section 2, the proposed hybrid compensation technique is presented. In Section 3, some considerations about the feature extraction are included. An overview of MEMLIN with cross-probability model based on GMMs is included in Section 4. The considered acoustic modelling in this work is presented in Section 5. In Section 6, the results with Hiwire project database are included. Finally, the conclusions are presented in Section 7.

2. The proposed system

The scheme of the proposed system for the non-native adaptation task is depicted in Fig. 1. In this task, stereo clean and noisy data are available, so that the proposed solution is composed of two phases: training and decoding. In the training phase, the available clean and noisy training feature vectors for each speaker and noise condition (“HTK feature extraction”) are needed to estimate the MEMLIN parameters (“MEMLIN training”) [3]. Furthermore the noisy training feature vectors are normalized using MEMLIN (“Normalization MEMLIN”), and matched acoustic models for each speaker and noise con-

This work has been supported by the national project TIN 2005-08660-C04-01

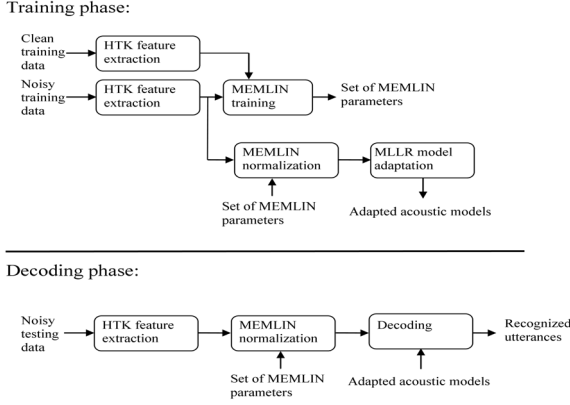


Figure 1: Scheme of the proposed system.

dition are obtained using the MLLR method (“MLLR model adaptation”). On the other hand, in the decoding phase, the MEMLIN-normalized testing feature vectors (“HTK feature extraction” and “Normalization MEMLIN”) are recognized with the corresponding adapted acoustic models (“Decoding”).

3. Feature vector extraction

In this work, the static coefficients are composed by 13 Mel Frequency Cepstral coefficients, including C_0 th coefficient, which are computed with the HTK toolkit [6]. MEMLIN is applied only to the static coefficients. Posteriorly, the normalized feature vector is completed with 3 derivatives which are calculated with three linear projections of length 9 frames.

4. Feature vector normalization

MEMLIN is an empirical feature vector normalization technique which uses stereo data in order to estimate the different compensation linear transformations in a previous training process. The clean feature space is modelled as a mixture of Gaussians. The noisy space is split into several basic acoustic environments and each one is modelled as a mixture of Gaussians. The linear transformations are estimated for all basic environments between a clean Gaussian and a noisy Gaussian. A scheme of MEMLIN can be shown in Fig. 2.

4.1. MEMLIN approximations

• Clean feature vectors, \mathbf{x}_t , are modelled using a GMM of C components

$$p(\mathbf{x}_t) = \sum_{s_x=1}^C p(\mathbf{x}_t|s_x)p(s_x), \quad (1)$$

$$p(\mathbf{x}_t|s_x) = \mathcal{N}(\mathbf{x}_t; \mu_{s_x}, \Sigma_{s_x}), \quad (2)$$

where μ_{s_x} , Σ_{s_x} , and $p(s_x)$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with the clean model Gaussian s_x .

• Noisy space is split into several basic environments, e , and the noisy feature vectors, \mathbf{y}_t , are modeled as a GMM of C' components for each basic environment

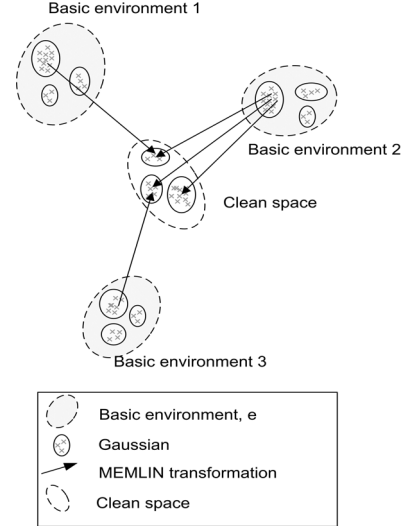


Figure 2: Scheme of MEMLIN.

$$p_e(\mathbf{y}_t) = \sum_{s_y^e=1}^{C'} p(\mathbf{y}_t|s_y^e)p(s_y^e), \quad (3)$$

$$p(\mathbf{y}_t|s_y^e) = \mathcal{N}(\mathbf{y}_t; \mu_{s_y^e}, \Sigma_{s_y^e}), \quad (4)$$

where s_y^e denotes the corresponding Gaussian of the noisy model for the e basic environment; $\mu_{s_y^e}$, $\Sigma_{s_y^e}$, and $p(s_y^e)$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with s_y^e .

• Clean feature vectors can be approximated as a linear function of the noisy feature vector, which depends on the basic environment and the clean and noisy model Gaussians: $\mathbf{x} \approx \Psi(\mathbf{y}_t, s_x, s_y^e) = \mathbf{y}_t - \mathbf{r}_{s_x, s_y^e}$, where \mathbf{r}_{s_x, s_y^e} is a bias vector transformation between noisy and clean feature vectors for each pair of Gaussians, s_x and s_y^e .

4.2. MEMLIN enhancement

With those approximations, MEMLIN transforms the MMSE estimation expression, $\hat{\mathbf{x}}_t = E[\mathbf{x}|\mathbf{y}_t]$, into

$$\hat{\mathbf{x}}_t = \mathbf{y}_t - \sum_e \sum_{s_y^e} \sum_{s_x} \mathbf{r}_{s_x, s_y^e} p(e|\mathbf{y}_t) p(s_y^e|\mathbf{y}_t, e) p(s_x|\mathbf{y}_t, e, s_y^e), \quad (5)$$

where $p(e|\mathbf{y}_t)$ is the a posteriori probability of the basic environment; $p(s_y^e|\mathbf{y}_t, e)$ is the a posteriori probability of the noisy model Gaussian, s_y^e , given the feature vector, \mathbf{y}_t , and the basic environment, e . To estimate those terms, $p(e|\mathbf{y}_t)$ and $p(s_y^e|\mathbf{y}_t, e)$, equations (3) and (4) are applied as described in [7]. Finally, the cross-probability model, $p(s_x|\mathbf{y}_t, e, s_y^e)$, is the probability of the clean model Gaussian, s_x , given the feature vector, \mathbf{y}_t , the basic environment, e , and the noisy model Gaussian, s_y^e . The bias vector transformation, \mathbf{r}_{s_x, s_y^e} , is estimated in a training phase using stereo data [7], while $p(s_x|\mathbf{y}_t, e, s_y^e)$ is modeled as with a GMM.

4.3. Cross-probability model based on GMM

The noisy feature vectors associated to a pair of Gaussians (s_x and s_y^e) are modeled with a GMM of C'' components

	# Sent	Corr	Sub	Del	Ins	Err	S. Err
Sum/Avg NNA LN MLLR	4038	71.67	16.13	12.21	0.33	28.67	41.08
Sum/Avg NNA MN MLLR	4038	45.36	25.26	29.38	0.28	54.92	66.52
Sum/Avg NNA HN MLLR	4038	6.34	32.65	61.00	0.08	93.74	94.82

Table 1: Average baseline results for noisy tests with simultaneous speaker and noise adaptation using an adaptation set of 50 utterances with MLLR, where LN, MN and HN represents Low, Medium and High Noise, respectively.

$$p(\mathbf{y}_t | s_x, s_y^e) = \sum_{s'_y=1}^{C''} p(\mathbf{y}_t | s_x, s_y^e, s'_y) p(s'_y | s_x, s_y^e), \quad (6)$$

$$p(\mathbf{y}_t | s_x, s_y^e, s'_y) = \mathcal{N}(\mathbf{y}_t; \mu_{s_x, s_y^e, s'_y}, \Sigma_{s_x, s_y^e, s'_y}), \quad (7)$$

where μ_{s_x, s_y^e, s'_y} , $\Sigma_{s_x, s_y^e, s'_y}$, and $p(s'_y | s_x, s_y^e)$ are the mean, the diagonal covariance matrix, and the a priori probability associated with s'_y . Gaussian of the cross-probability GMM associated with s_x and s_y^e . To train these three parameters, the EM algorithm is applied [3]. So, $p(s_x | \mathbf{y}_t, e, s_y^e)$ can be obtained with (6) as

$$p(s_x | \mathbf{y}_t, e, s_y^e) = \frac{p(\mathbf{y}_t | s_x, s_y^e) p(s_x, s_y^e)}{\sum_{s_x} p(\mathbf{y}_t | s_x, s_y^e) p(s_x, s_y^e)}, \quad (8)$$

where $p(s_x, s_y^e)$ is the probability of the Gaussians s_x and s_y^e , which is estimated with EM algorithm [3].

For the work reported in this paper, the MEMLIN parameters and the cross-probability models for each speaker and noise condition were trained using the identical 50 utterances for the clean adaptation set and the noisy adaptation set. Thus, each testing utterance is compensated only with the corresponding MEMLIN parameters and the cross-probability model obtained with the same kind of noise and speaker, tuning our cepstral enhancement technique on the noise types and speakers ($\#e = 1$). 128 Gaussians per noisy ($C' = 128$) and clean ($C = 128$) spaces are used. Finally the noisy feature vectors associated to a pair of Gaussians are modeled with a GMM of 2 ($C'' = 2$) components.

5. Acoustic modelling

For the acoustic modelling, the HTK toolkit [6] was used to obtain the models used for the baseline. The initial clean user-independent acoustic models were obtained with the TIMIT database, where the whole training set of TIMIT was used, which means a total of 5376 utterances for training. A set of 46 acoustic models were used (each one of them representing a phoneme linguistic unit). Each unit was modelled with 3 states; and every state was modelled with a Gaussian Mixture Model containing 128 Gaussians per state.

6. Non-native adaptation task (NNA) results

The clean set of the HIWIRE project database contains 8100 English utterances pronounced by non-native speakers (French, Greek, Italian and Spanish). The collect utterances correspond to command and control aeronautics application. To build the noisy sets, noise recorded in an airplane cockpit was artificially

added to the clean data, defining three noise conditions according the SNR: low, medium and high noise.

In the defined non-native adaptation task, the corpus is split into two sets: adaptation and testing. The adaptation one is composed of 50% utterances of each speaker and includes stereo clean and noisy signals. The purpose of this task is the evaluation of acoustic model adaptation and feature vector normalization methods.

6.1. Noisy tests with simultaneous speaker and noise adaptation: baseline results

As baseline results for the non-native adaptation task, MLLR algorithm is considered. Thus, the acoustic models are adapted to simultaneous speaker and noise adaptation using 50 utterances per speaker and noise condition and a 32 class regression tree is applied. The average results are presented in Table 1 for low, medium and high noise conditions (“Sum/Avg NNA LN MLLR”, “Sum/Avg NNA MN MLLR” and “Sum/Avg NNA HN MLLR”, respectively). It can be observed the effect of MLLR, which reduces significantly the average WER in all conditions (if no adaptation technique is applied, the average WER for low, medium and high noise conditions are 54.92%, 77.17% and 97.88%, respectively when all the testing utterances for each noise condition are recognized, 8081).

6.2. Noisy tests with simultaneous speaker and noise adaptation results: the proposed system

Tables 2, 3 and 4 show the performance of the proposed system for low, medium and high noise conditions, respectively. Since the MEMLIN parameters were trained on the same noise and speaker conditions of the testing utterances, the combination of MEMLIN and MLLR provides a reasonable performance improvement over the baseline, which consisted on MLLR, (6.54%, 13.52% and 53.45% of average WER for low, medium and high noise conditions, respectively). Note the importance of the feature vector adaptation, which generates a more homogeneous normalized space than the noisy one, being able to train more satisfactory MLLR-adapted acoustic models.

7. Conclusions

The results exposed in this work show the good performance obtained in the Hiwire project database by the combined use of a feature vector normalization technique (MEMLIN) and a model adaptation method (MLLR). The result in the non-native adaptation (NNA) task of a 6.54% of WER across the four non-native tasks (French, Greek, Italian and Spanish speakers) in the low noise environment is very fruitful; which gets reinforced with the encouraging result of a 13.52% of average WER of the medium noise environment and the 53.45% of WER in the high noise environment.

As a conclusion to this work, it is shown that both kind

Low Noise	# Sent	Corr	Sub	Del	Ins	Err	S. Err
French	1549	93.65	5.41	0.95	0.29	6.65	12.52
Greek	1000	93.59	4.99	1.42	0.33	6.74	12.10
Italian	990	94.29	4.95	0.76	0.90	6.61	12.93
Spanish	499	94.83	4.37	0.80	0.53	5.70	12.42
Sum/Avg	4038	93.94	5.06	1.00	0.48	6.54	12.51

Table 2: Low noise results for simultaneous speaker and noise adaptation for an adaptation set of 50 utterances with the proposed system.

Medium Noise	# Sent	Corr	Sub	Del	Ins	Err	S. Err
French	1549	85.76	11.53	2.70	0.52	14.76	23.11
Greek	1000	85.59	11.60	2.81	1.19	15.60	24.30
Italian	990	89.97	8.40	1.63	1.16	11.19	18.79
Spanish	499	90.06	8.88	1.06	0.46	10.40	18.44
Sum/Avg	4038	87.32	10.43	2.25	0.84	13.52	21.77

Table 3: Medium noise results for simultaneous speaker and noise adaptation for an adaptation set of 50 utterances with the proposed system.

High Noise	# Sent	Corr	Sub	Del	Ins	Err	S. Err
French	1549	45.53	40.57	13.90	3.49	57.96	66.43
Greek	1000	46.63	43.62	9.75	7.34	60.71	69.30
Italian	990	57.77	31.37	10.86	3.72	45.95	53.43
Spanish	499	61.63	32.21	6.16	2.25	40.62	51.30
Sum/Avg	4038	50.91	37.98	11.11	4.36	53.45	62.09

Table 4: High noise results for simultaneous speaker and noise adaptation for an adaptation set of 50 utterances with the proposed system.

of techniques (feature vector normalization and model adaptation) can be very useful working together in an additive noise environment with a wide set of non-native speakers. A future work in this direction should be the testing of the combination of these techniques in a real noise environment to know the jointly effect of additive noise and convolutional distortion in the proposed hybrid system performance.

8. References

- [1] L. Neumeyer and M. Weintraub, "Robust Speech Recognition in Noise Using Adaptation and Mapping Techniques," in *Proceedings of ICASSP*, vol. 1, 1995, pp. 141–144.
- [2] A. Sankar and C. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 190–202, May 1996.
- [3] L. Buera, E. Lleida, J. Nolzco, A. Miguel, and A. Ortega, "Time-dependent cross-probability model for multi-environment model based linear normalization," in *ICSLP*, Sept. 2006.
- [4] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 291–298, Apr 1994.
- [5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.
- [6] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The htk book (for htk version 3.3)," 1995-1999 Microsoft Corporation, 2001-2005 Cambridge University Engineering Department, Tech. Rep.
- [7] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz, "Cepstral vector normalization based on stereo data for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 15, pp. 1098–1113, March 2007.