ROBUST SPEECH RECOGNITION WITH ON-LINE UNSUPERVISED ACOUSTIC FEATURE COMPENSATION

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

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

ABSTRACT

An on-line unsupervised hybrid compensation technique is proposed to reduce the mismatch between training and testing conditions. It combines Multi-Environment Model based LInear Normalization with cross-probability model based on GMMs (MEMLIN CPM) with a novel acoustic model adaptation method based on rotation transformations. Hence, a set of rotation transformations is estimated with clean and MEMLIN CPM-normalized training data by linear regression in an unsupervised process. Thus, in testing, each MEMLIN CPM normalized frame is decoded using a modified Viterbi algorithm and expanded acoustic models, which are obtained from the reference ones and the set of rotation transformations. To test the proposed solution, some experiments with Spanish SpeechDat Car database were carried out. MEMLIN CPM over standard ETSI front-end parameters reaches 83.89% of average improvement in WER, while the introduced hybrid solution goes up to 92.07%. Also, the proposed hybrid technique was tested with Aurora 2 database, obtaining an average improvement of 68.88% with clean training.

Index Terms— robust speech recognition, feature vector normalization, acoustic model adaptation.

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 two main lines of research: acoustic model adaptation methods, and feature vector normalization methods. Also hybrid techniques, which are the combination of a feature vector normalization method and an acoustic model adaptation method, exist and they have proved to be effective [1].

In general, acoustic model adaptation methods, e.g. Maximum A Posteriori, MAP, [2], Maximum Likelihood Linear Regression, MLLR, [3]..., produce better results [4] because they can model the uncertainty caused by the noise statistics by mapping the parameters of the reference acoustic models to the noisy space. Thus, acoustic model methods take into account implicitly all kinds of degradations of the feature vectors. However, these methods usually require more data and computing time than feature vector normalization methods do and their performances degrade when the transcription of the adaptation data is not available (unsupervised methods) [5].

On the other hand, feature vector normalization methods, e. g. multivariate Gaussian-based cepstral normalization, RATZ, [6], or Stereo based Piecewise LInear Compensation for Environments, SPLICE, [7], which can be grouped into different classes [8], provide more on-line solutions than acoustic model adaptation methods using, in general, less adaptation data.

A previous work [9] shows that Multi-Environment Model-based LInear Normalization with cross-probability model based on GMMs, MEMLIN CPM, (an empirical feature vector normalization method based on stereo data and the MMSE estimator) is effective to compensate the effects of dynamic and adverse car conditions, improving the performance of techniques based on similar criterions, e.g. RATZ or SPLICE. However, all these techniques estimate the clean feature vector by using a bias vector transformation, not taking into account several kinds of degradation, like rotations or variance deformations. To compensate these effects, in this work we propose an on-line unsupervised hybrid solution which combines MEMLIN with cross-probability model based on GMMs with a novel acoustic model adaptation method based on rotation transformations over an expanded HMM-state space.

This paper is organized as follows: In Section 2, the proposed hybrid compensation technique is presented. In Section 3, some considerations about MEMLIN CPM are included. The rotation matrix estimation process is explained in Section 4. The on-line selection of the rotation matrix for each normalized feature vector in the decoding process is presented in Section 5. In Section 6, the results with Spanish SpeechDat Car [10] and Aurora 2 [11] databases are included. Finally, the conclusions and future lines are presented in Section 7.

2. UNSUPERVISED HYBRID COMPENSATION

The scheme of the proposed unsupervised hybrid compensation technique is depicted in Fig. 1. It is composed of two phases: training and decoding. In the unsupervised train-

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

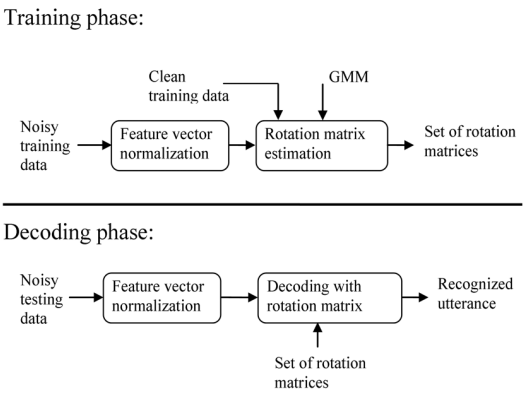

Fig. 1. Scheme of the proposed unsupervised hybrid compensation technique.

ing phase, the noisy training data are compensated with the corresponding feature vector normalization method, "Feature vector normalization", (MEMLIN CPM in our case), and the clean and normalized spaces are modelled with GMMs, "GMM". Also, a set of rotation matrices is estimated by linear regression with the normalized and clean stereo training data ("Rotation matrix estimation"), obtaining one rotation matrix for each pair of Gaussians (clean-normalized). On the other hand, in the decoding phase, each normalized testing feature vector ("Feature vector normalization") is recognized with expanded acoustic models ("Decoding with rotation matrix"), which are obtained with the reference acoustic models and a selected rotation matrix. The selected rotation matrix is obtained implicity during the search process from the associated expanded state by using the ML criterion in a modified Viterbi algorithm. Note that there is not any restriction about the feature vector normalization method, so that anyone can be used in this scheme.

3. FEATURE VECTOR NORMALIZATION

MEMLIN CPM [9] is the on-line selected feature vector normalization technique for the hybrid compensation method in this work, although other algorithms could be used. MEMLIN CPM is based on three approximations: the clean feature space is modelled as a GMM; the noisy space is split into several basic acoustic environments and each one of them is modelled as a GMM. Finally, the third assumption consists on defining a bias vector transformation associated with each pair of Gaussians from the clean and the noisy basic environment spaces. To compensate a testing feature vector, the MMSE estimator is used, where the cross-probability model (the probability of the clean model Gaussian given the noisy model Gaussian and the noisy feature vector), which has a relevant importance, is modelled with GMMs as [9].

It can be observed that the clean estimated feature vector that MEMLIN CPM provides for the time index t, $\hat{\mathbf{x}}_t$, is a shifted version of the noisy one $\mathbf{y}_t$: $\hat{\mathbf{x}}_t = \mathbf{y}_t + \mathbf{g}_t$, where $\mathbf{g}_t$ is the corresponding bias vector which depends on the acoustic environment and the noisy and clean GMM modelled spaces.

Note that MEMLIN CPM can be seen as an acoustic model adaptation technique if the reference acoustic models are composed by HMMs with GMMs as pdfs for the different states. Thus, the score of the normalized feature vector, $\hat{\mathbf{x}}_t$, given a Gaussian, $\mathcal{N}(\hat{\mathbf{x}}_t, \mu_{ref}, \mathbf{\Sigma}_{ref})$, is the same that the score of the noisy feature vector, $\mathbf{y}_t$, given an adapted Gaussian, $\mathcal{N}(\mathbf{y}_t, \mu_{ref} - \mathbf{g}_t, \mathbf{\Sigma}_{ref})$. Furthermore, all the feature vector normalization methods which transform the noisy feature vector using only a bias vector (e. g. Cepstral Mean Normalization (CMN), RATZ, SPLICE, MEMLIN CPM...) can be seen also as acoustic model adaptation techniques which transform just the mean vectors each time index.

4. ROTATION MATRIX ESTIMATION

In order to define a set of rotation matrices which determines the relation between clean and normalized feature vectors, three approximations are considered

- Clean feature vectors, $\mathbf{x}_t$, are modelled using a GMM

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

$$p(\mathbf{x}_t|s_x) = \mathcal{N}(\mathbf{x}_t; \mu_{s_x}, \mathbf{\Sigma}_{s_x}), \tag{2}$$

  where μ_{s_x}, $\mathbf{\Sigma}_{s_x}$, 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.

- Normalized feature vectors, $\hat{\mathbf{x}}_t$, are modelled using a GMM

$$p(\hat{\mathbf{x}}_t) = \sum_{s_{\hat{x}}} p(\hat{\mathbf{x}}_t|s_{\hat{x}})p(s_{\hat{x}}), \tag{3}$$

$$p(\hat{\mathbf{x}}_t|s_{\hat{x}}) = \mathcal{N}(\hat{\mathbf{x}}_t; \mu_{s_{\hat{x}}}, \mathbf{\Sigma}_{s_{\hat{x}}}), \tag{4}$$

  being $\mu_{s_{\hat{x}}}$, $\mathbf{\Sigma}_{s_{\hat{x}}}$, and $p(s_{\hat{x}})$ the mean vector, the diagonal covariance matrix, and the a priori probability associated with the normalized model Gaussian $s_{\hat{x}}$.

- Normalized feature vectors can be approximated as a linear function of the clean feature vectors which depends on the clean and normalized model Gaussians s_x and $s_{\hat{x}}$: $\hat{\mathbf{x}}_t \approx \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t$, where $\mathbf{A}_{s_x,s_{\hat{x}}}$ is the rotation matrix between $\hat{\mathbf{x}}_t$ and $\mathbf{x}_t$ associated to the pair of Gaussians s_x and $s_{\hat{x}}$.

Hence, a set of rotation matrices can be defined as

$$\mathcal{A} = \{\mathbf{A}_{s_x,s_{\hat{x}}}\}_{s_x=1,s_{\hat{x}}=1}^{\#s_x,\#s_{\hat{x}}} = \{\mathbf{A}_n\}_{n=1}^{N}, \tag{5}$$

where, to simplify the notation, the index n represents each pair of Gaussians s_x, $s_{\hat{x}}$ and N denotes the number of the pair of Gaussians: $N = \#s_x \times \#s_{\hat{x}}$.

To estimate the rotation matrices $\mathbf{A}_n$, clean and normalized stereo data are used in the previous unsupervised training phase: $(\mathbf{X}^{Tr}, \hat{\mathbf{X}}^{Tr}) =$

$$\xi_n = \frac{1}{T}\sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr}) \cdot Tra\left[(\hat{\mathbf{x}}_t^{Tr} - \mathbf{A}_n\mathbf{x}_t^{Tr})(\hat{\mathbf{x}}_t^{Tr} - \mathbf{A}_n\mathbf{x}_t^{Tr})^T\right]. \quad (6)$$

$$\mathbf{A}_n = \mathbf{A}_{s_x,s_{\hat{x}}} = \underset{\mathbf{A}_n}{arg\,min}\{\xi_n\} = \left[\sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})(\hat{\mathbf{x}}_t^{Tr}(\mathbf{x}_t^{Tr})^T)\right] \cdot \left[\sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})(\mathbf{x}_t^{Tr}(\mathbf{x}_t^{Tr})^T)\right]^{-1}. \quad (7)$$

$\{(\mathbf{x}_1^{Tr}, \hat{\mathbf{x}}_1^{Tr}); ...; (\mathbf{x}_t^{Tr}, \hat{\mathbf{x}}_t^{Tr}); ...; (\mathbf{x}_T^{Tr}, \hat{\mathbf{x}}_T^{Tr})\}$, with $t \in [1, T]$, where $\hat{\mathbf{X}}^{Tr}$ is obtained after normalizing the noisy training data $\mathbf{Y}^{Tr}$ (MEMLIN CPM has been selected for this purpose in this work). Thus, $\mathbf{A}_n$ is obtained by minimizing the defined mean weighted square error, ξ_n, (6) with respect to $\mathbf{A}_n$ (7), where $Tra[\bullet]$ is the trace, $(\bullet)^T$ is the transpose, $p(s_x|\mathbf{x}_t^{Tr})$ is the a posteriori probability of the clean model Gaussian s_x, given the clean training feature vector $\mathbf{x}_t^{Tr}$, and $p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})$ is the a posteriori probability of the normalized model Gaussian $s_{\hat{x}}$, given the normalized training feature vector $\hat{\mathbf{x}}_t^{Tr}$.

$$p(s_x|\mathbf{x}_t^{Tr}) = \frac{p(\mathbf{x}_t^{Tr}|s_x)p(s_x)}{\sum_{s_x} p(\mathbf{x}_t^{Tr}|s_x)p(s_x)}, \quad (8)$$

$$p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr}) = \frac{p(\hat{\mathbf{x}}_t^{Tr}|s_{\hat{x}})p(s_{\hat{x}})}{\sum_{s_{\hat{x}}} p(\hat{\mathbf{x}}_t^{Tr}|s_{\hat{x}})p(s_{\hat{x}})}. \quad (9)$$

5. DECODING WITH ROTATION MATRIX SELECTION

Given a normalized feature vector, $\hat{\mathbf{x}}_t$, a rotation matrix, $\mathbf{A}_t$, is selected from the set of estimated rotation matrices, $\mathbf{A}_n$, by ML maximization criterion in a similar way as [12]. Hence, a set of expanded acoustic models is built, where each q state of the clean space HMM acoustic models, ($q \in [1, Q]$), is expanded into N states (q, n) considering the linear approximation $\hat{\mathbf{x}}_t \approx \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t = \mathbf{A}_n\mathbf{x}_t$. The goal of the state expansion is to reduce the mismatch between the clean space acoustic models and the normalized feature vectors for each rotation transformation. Thus, each expanded state is specialized in one of the rotation transformations. Assuming that a component s_q in the pdf mixture of the original state q follows a normal distribution: $\mathcal{N}(\mathbf{x}_t; \mu_{s_q}, \mathbf{\Sigma}_{s_q})$, the corresponding expanded component $s_{q,n}$ is assumed to follow the distribution $\mathcal{N}(\hat{\mathbf{x}}_t; \mathbf{A}_n\mu_{s_q}, \mathbf{A}_n\mathbf{\Sigma}_{s_q}\mathbf{A}_n^T)$. So, finally the pdf for the expanded state (q, n), $p(\hat{\mathbf{x}}_t|q, n)$, is a GMM composed by the defined expanded components, where the a priori component weights remain unaltered: $p(s_{q,n}) = p(s_q)$:

$$p(\hat{\mathbf{x}}_t|q, n) = \sum_{s_q} p(s_q)\mathcal{N}(\hat{\mathbf{x}}_t; \mathbf{A}_n\mu_{s_q}, \mathbf{A}_n\mathbf{\Sigma}_{s_q}\mathbf{A}_n^T). \quad (10)$$

Note that the proposed expanded acoustic models, from a generative point of view, can be seen as a more flexible speech production process for normalized space, since they can generate sequences of rotated feature vectors more suitable to the normalized space.

Once the reference acoustic models have been expanded, the classic search algorithm (Viterbi) for decoding unlabeled sequences has to be modified in a similar way as [13] [14], computing recursively the score state variable, $\phi_{q,n}(t)$, for the state (q, n) and the time index t (11). In this way, the rotation matrix $\mathbf{A}_t$ for each normalized feature vector is determined implicity for the sequence of expanded states which maximizes the likelihood at the end of the utterance.

$$\phi_{q,n}(t) = \max_{q',n'}\{\phi_{q',n'}(t-1) \cdot \pi_{q',n',q,n} \cdot p(\hat{\mathbf{x}}_t|q, n)\}, \quad (11)$$

being $\pi_{q',n',q,n}$ the transition probability from expanded state (q, n) to (q', n'), which in this work is assumed to be

$$\pi_{q',n',q,n} \approx \pi_{q',q}\pi_{n',n} \approx \frac{\pi_{q',q}}{N}, \quad (12)$$

where $\pi_{q',q}$ is the transition probability from the reference state q to q', and $\pi_{n',n}$ is the transition probability from the matrix $\mathbf{A}_n$ to $\mathbf{A}_{n'}$, which is considered equiprobability for all transitions.

Note that the selection of $\mathbf{A}_t$ is made at the same time of decoding using the expanded acoustic models and the modified Viterbi algorithm. Thus, the presented hybrid solution can be seen as decoding each MEMLIN CPM normalized feature vector, $\hat{\mathbf{x}}_t = \mathbf{y}_t + \mathbf{g}_t$, with the corresponding expanded acoustic models, where the mean vectors and covariance matrices are $\mathbf{A}_t\mu$ and $\mathbf{A}_t\mathbf{\Sigma}\mathbf{A}_t^T$, respectively. Note that this solution provides the same results that recognizing the noisy feature vector, $\mathbf{y}_t$, with new acoustic models where the mean vectors and covariance matrices are $\mathbf{A}_t\mu - \mathbf{g}_t$ and $\mathbf{A}_t\mathbf{\Sigma}\mathbf{A}_t^T$, respectively. This point of view is conceptually similar to MLLR, where shift and rotation are included in acoustic models. However, the shift and rotation transformations for the proposed hybrid technique are selected for each feature vector and are estimated with a different criterion than MLLR. Also, the unsupervised MLLR version needs a previous step to provide an estimation of the transcription of the adaptation data (usually a recognition process), so that the performance of the unsupervised MLLR solution can degrade dramatically, especially in high noise conditions or difficult tasks (e.g. large vocabulary, spontaneous speech...) due to the estimation of the transcription can not be precise enough. These problems do not affect to the proposed hybrid on-line technique, which is unsupervised and does not precise the transcription of the adaptation data.

Train	Test	E1	E2	E3	E4	E5	E6	E7	AWER (%)
CLK	CLK	0.95	2.32	0.70	0.25	0.57	0.32	0.00	0.91
CLK	HF	3.05	13.29	15.52	27.32	31.36	35.56	53.06	21.48
HF	HF	3.81	6.86	3.50	3.76	4.96	4.44	3.06	4.63
† HF	HF	1.14	4.37	1.68	2.13	2.10	2.06	23.13	3.42
HF MLLR	HF	1.33	4.55	2.52	3.63	7.34	5.24	26.19	5.28
CLK A	HF MEMLIN CPM	2.10	3.86	1.54	2.88	2.86	1.43	1.70	2.54

Table 1. WER baseline results, in %, from the different basic environments (E1,..., E7) of Spanish SpeechDat Car database, where MWER is the Mean WER.

6. RESULTS

To study the performance of the proposed unsupervised online hybrid compensation technique, a set of experiments were carried out using two databases. On one hand, the Spanish SpeechDat Car database [10], which is composed by real, dynamic, and complex environments. On the other hand, Aurora 2 [11], which does not represent real environments due to the noise has been artificially added, but it has been widely applied to compare robustness techniques.

In both cases, the recognition task is isolated and continuous digits recognition. As feature set, the standard ETSI front-end [15] features plus energy and the corresponding delta and delta delta coefficients are used. Cepstral mean normalization is applied to testing and training data. The different feature vector normalization techniques are applied to the 12 MFCCs and energy, whereas the derivatives are computed over the normalized static coefficients. The acoustic models are composed of 16 state HMM for each digit, a 3 state begin-end silence HMM and a 1 state inter-word silence HMM. In all cases, each pdf state is composed by a mixture of three Gaussians.

6.1. Results with Spanish SpeechDat Car corpus

Seven basic environments were defined: car stopped, motor running (E1), town traffic, windows close and climatizer off (silent conditions) (E2), town traffic and noisy conditions: windows open and/or climatizer on (E3), low speed, rough road, and silent conditions (E4), low speed, rough road, and noisy conditions (E5), high speed, good road, and silent conditions (E6), and high speed, good road, and noisy conditions (E7).

The clean signals are recorded with a CLose talK (CLK) microphone (Shune SM-10A), and the noisy ones are recorded by a Hands-Free (HF) microphone placed on the ceiling in front of the driver (Peiker ME15/V520-1). The SNR range for CLK signals goes from 20 to 30 dB, and for HF ones goes from 5 to 20 dB. The unsupervised training process has been carried out with CLK and HF signals of the training set.

The Word Error Rate (WER) baseline results for each basic environment are presented in Table 1, where AWER is the Average WER computed proportionally to the number of utterances in each basic environment. "Train" column refers to the signals used to obtain the corresponding acoustic HMMs: CLK if they are trained with all clean training utterances, and HF and if they are trained with all noisy ones. † HF indicates that specific acoustic models are trained for each basic environment. All acoustic models are obtained with ML algorithm. "Test" column indicates which signals are used for recognition: clean, CLK, or noisy, HF.

Table 1 shows the effect of real car conditions, which increases the WER in all of the basic environments, (Train CLK, Test HF), concerning the rates for clean conditions, (Train CLK, Test CLK). When acoustic models are retrained with ML algorithm using all basic environment signals (Train HF), AWER decreases, 4.63%. Finally, the most competitive results (3.42% AWER) are obtained when specific acoustic models are retrained for each basic environment with ML algorithm, (Train † HF), despite the poor WER reached with E7 due to the reduced amount of data for that condition (67 utterances). However, this option is not possible in a real situation because the basic environment can not be known for each testing utterance.

Figure 2 shows the Average Improvement in WER (AIMP) in % for MEMLIN CPM and the proposed hybrid technique based on MEMLIN CPM (MEMLIN CPM A) when different number of Gaussians per basic environment are considered for the feature vector normalization techniques (4, 8, 16, 32, 64 and 128). Furthermore, SPLICE with Environmental Model selection (SPLICE EM) [16]) is included to compare. In case of MEMLIN CPM, clean feature space is modelled with the same number of Gaussians than the basic environments and the cross-probability model is composed by 2 Gaussians. Also, 17 rotation matrices are estimated in all cases ($\#s_x = \#s_{\hat{x}} = 4$, plus the identity matrix). AIMP is computed with AWER as

$$AIMP = \frac{100(AWER - AWER_{CLK-HF})}{AWER_{CLK-CLK} - AWER_{CLK-HF}}, \quad (13)$$

where $AWER_{CLK-CLK}$ is the mean WER obtained with clean conditions (0.91 in this case), and $AWER_{CLK-HF}$ is the baseline (21.48). So, A 100% AIMP would be achieved

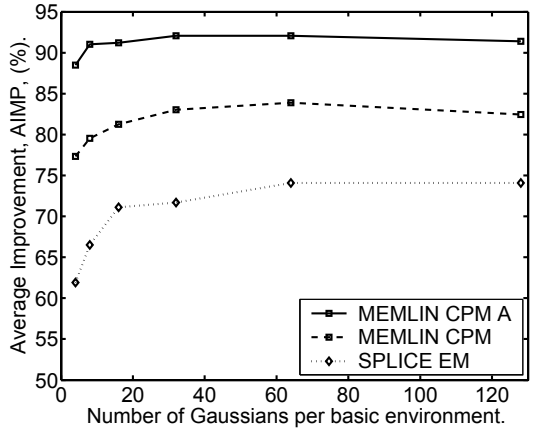

Fig. 2. Average improvement in WER, AIMP, in % with Spanish SpeechDat Car database for different normalization techniques: SPLICE with environmental model selection (SPLICE EM), MEMLIN with Cross-Probability Model based on GMMs (MEMLIN CPM) and the proposed hybrid technique based on MEMLIN CPM and acoustic model adaptation based on rotation transformations (MEMLIN CPM A).

when AWER equals the one obtained under clean conditions.

It can be verified in Fig. 2 the important improvement that the presented hybrid solution obtains when it is applied over MEMLIN CPM for any number of Gaussians per basic environment concerning SPLICE EM and MEMLIN CPM. In fact, the performance with 32 components per basic environment (92.07% AIMP, 2.54% AWER) is significantly better than the best results for SPLICE EM (74.08% AIMP, 6.25% AWER) and MEMLIN CPM (83.89% AIMP, 4.23% AWER); even if matched training condition (81.93% AIMP, 4.63% AWER) or specific acoustic models for each basic environment (87.81% AIMP, 3.42% AWER) are considered, the performances are slightly inferior with respect to the one obtained with the proposed hybrid solution due to the noisy space is more heterogenous than the normalized one. Also, note that a reduced number of Gaussians per environment is enough to obtain satisfactory results (88.49% AIMP, 3.28% AWER with only two components per basic environment). The complete best WER results obtained with the hybrid solution are also included in Table 1 (Train CLK A, Test HF MEMLIN). Also the performance of unsupervised MLLR, where the transcription obtained from the decoding of the noisy data is assumed as the true one, is presented in Table 1 to complete the comparison (Train HF MLLR, Test HF). Note that the reached performance in this case (AWER 5.28%, 78.77% AIMP) is inferior than the match training condition results and the ones obtained with the proposed hybrid technique.

6.2. Results with Aurora 2 corpus

For the Aurora work, identical utterances from the clean training set and the multicondition training set have been used in the unsupervised training process for the hybrid solution. Thus, the noise types from set B and C keep as unseen conditions, while the system is tuned on the noise types from set A. All the improvements are computed with respect to the results reached with standard ETSI front-end (58.06% of average).

Figure 3 shows the word accuracy results in % with clean training for the proposed hybrid technique based on MEMLIN CPM when basic environments and clean space are modelled with 128 Gaussians. The cross-probability model is composed by 2 Gaussians and 17 rotation matrices are estimated ($\#s_x = \#s_{\hat{x}} = 4$) plus the identity matrix.

It can be observed an important improvement in set A, 71.99%, where the proposed hybrid technique is applied over the same kinds of noise which have been observed in the training process. Also competitive results have been reached in set B (73.31% of average improvement), where the testing conditions include different kinds of additive noise that the ones considered in the training process. However, the results are not as satisfactory for set C, which includes a different convolutional distortion from the training set. This indicates that the transformations learnt with the training data may be not representative of those required in set C. In summary, from the results in Fig. 3, a reasonable and consistent improvement for all the noise conditions can be appreciated, obtaining an average improvement slightly better than the one obtained with ETSI Advanced front-end [17] (67.41%). Note that, comparing with ETSI Advanced front-end results, the behavior of the proposed hybrid technique under seen conditions (set A) is much better (71.99% versus 66.57%), while for set B, which includes unseen additive noises, the average improvements are similar. On the other hand, the behavior of the proposed hybrid technique degrades in set C. From these results, we can conclude that a reasonable future work line could be to improve the performance of the presented technique under unseen conditions.

7. CONCLUSIONS

In this paper we have presented an on-line unsupervised hybrid compensation solution which combines Multi-Environment Model based LInear Normalization with cross-probability model based on GMMs, MEMLIN CPM, with a novel acoustic model adaptation technique based on rotation transformations which depend on GMMs. Although, other feature vector techniques can be used. The purpose of the hybrid solution is to compensate jointly the shift and rotation introduced by the acoustic environment. Some results with Spanish SpeechDat Car database show the effective performance of the proposed technique (92.07% of mean improvement with 32 Gaussians per basic environment) with respect to classic feature vector normalization techniques: SPLICE EM (74.08%), and MEMLIN CPM (83.89%), or acoustic model adaptation techniques: unsupervised MLLR (78.77%). Also important improvements have been obtained with Aurora 2 database (68.55%), even better than the one obtained with ETSI Advanced front-end (67.41%). As future lines we propose to use the hybrid solution with other feature vector

Clean training, multicondition testing															
	A					B					C				Percentage
	Subway	Babble	Car	Exhibition	Average	Restauran	Street	Airport	Station	Average	Subway M	Street M	Average	Average	Improvement
Clean	99,11	99,03	99,14	99,29	99,14	99,11	99,03	99,14	99,29	99,14	98,89	99,06	98,98	99,11	3,03%
20 dB	98,40	98,28	98,51	98,24	98,36	98,37	98,01	98,22	98,64	98,31	97,70	97,67	97,69	98,20	64,45%
15 dB	97,73	97,55	97,44	97,29	97,50	96,91	97,17	96,55	97,33	96,99	96,29	95,61	95,95	96,99	78,46%
10 dB	95,56	93,58	95,09	95,32	94,89	93,28	91,61	92,56	94,26	92,93	90,17	89,98	90,07	93,14	79,79%
5 dB	90,23	81,12	85,09	88,90	86,33	82,53	78,41	81,91	82,99	81,46	70,18	77,46	73,82	81,88	70,68%
0 dB	70,53	53,49	60,09	70,05	63,54	57,10	54,55	60,89	57,86	57,60	38,11	49,70	43,90	57,24	49,38%
-5dB	37,10	27,56	27,72	38,59	32,74	29,21	28,29	30,73	28,82	29,26	17,39	24,65	21,02	29,01	23,50%
Average	90,49	84,80	87,24	89,96	**88,12**	85,64	83,95	86,02	86,22	**85,46**	78,49	82,08	**80,28**	85,49	
Improvement	69,88%	77,41%	69,37%	71,31%	71,99%	77,48%	62,89%	77,34%	75,53%	73,31%	52,25%	52,06%	52,16%		68,55%
ETSI Adv.	56,32%	77,13%	70,10%	62,72%	66,57%	74,24%	63,17%	79,30%	76,35%	73,27%	59,91%	54,84%	57,37%		67.41%

Fig. 3. Word accuracy and improvement obtained for Aurora 2 corpus with the proposed hybrid technique based on MEMLIN CPM and acoustic model adaptation based on rotation transformations, and with ETSI Advanced front-end.

normalization techniques, and in different tasks, e. g. large vocabulary. Furthermore, we are working to improve the method when it is applied under unseen conditions.

8. REFERENCES

[1] 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.

[2] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriory estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 291–298, Apr 1994.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continous-density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[4] Leonardo Neumeyer and Mitchel Weintraub, "Robust Speech Recognition in Noise Using Adaptation and Mapping Techniques," in *Proceedings of ICASSP* , 1995, vol. 1, pp. 141–144.

[5] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-based unsupervised mllr for speaker adaptation," in *ASR*, 2000, vol. 2, pp. 128–132.

[6] P. Moreno, *Speech recognition in noisy environments*, Ph.D. thesis, ECE Department, Carnegie-Mellon University, Apr. 1996.

[7] J. Droppo, L. Deng, and A. Acero, "Evaluation of the splice algorithm on the aurora2 database," in *in Proc. Eurospeech*, Sept. 2001, vol. 1.

[8] R. M. Stern, B. Raj, and P.J. Moreno, "Compensation for environmental degradation in automatic speech recognition," in *in Proc. of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-au-Mousson, France, April 1997, pp. 33–42.

[9] L. Buera, E. Lleida, J.A. Nolazco, A. Miguel, and A. Ortega, "Time-dependent cross-probability model for multi-environment model based linear normalization," in *ICSLP*, Sept. 2006.

[10] Henk van den Heuvel, Jerme Boudy, Robrecht Comeyne, Stephan Euler, Asuncion Moreno, and G. Richard, "The speechdat-car multilingual speech databases for in-car applications: some first validation results," in *Eurospeech*, 1999.

[11] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. in ISCA ITRW ASR2000*, Paris, France, September 2000.

[12] A. Miguel, E. Lleida, A. Juan, L. Buera, A. Ortega, and O. Saz, "Local transformation models for speech recognition," in *ICSLP*, Pittsburgh, USA, 2006.

[13] A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," in *Acoustics, Speech, and Signal Processing, ICASSP*, 1990, pp. 845–848.

[14] M. J. F. Gales and S. J. Young, "An improved approach to the hidden markov model decomposition of speech and noise," in *Proc. of ICASSP*, 1992, pp. 233–236.

[15] ETSI, "Speech processing transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," Tech. Rep., ETSI ES 201 108 version 1.1.2, April 2000.

[16] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the AURORA2 database," in *Eurospeech*, 2001.

[17] ETSI, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," Tech. Rep., ETSI ES 202 050 version 1.1.1, Oct. 2002.