On-Line Feature and Acoustic Model Space Compensation for Robust Speech Recognition in Car Environment

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

***Abstract*— In order to develop a robust man-machine interface based on speech for cars, the speaker variability and the acoustic environment effects have to be compensated. In this work, an on-line feature and acoustic model compensation (MATE-MEMLIN) is proposed to compensate the speaker variability and the acoustic car environment. MATE-MEMLIN consists on the combination of the techniques augMented stAte space acousTic modEl (MATE) and Multi-Environment Model based LInear Normalization (MEMLIN). MATE defines expanded acoustic models to compensate the speaker frequency variability using data driven estimated linear transformations. On the other hand, MEMLIN, an empirical feature vector normalization technique, was also presented and it was proved to be effective to compensate environment mismatch. Some experiments with Spanish SpeechDat Car database were carried out in order to study the performance of the proposed technique in a real car environment, reaching an important mean improvement in Word Error Rate, WER.**

I. INTRODUCTION

Since cars are more and more considered as business offices, drivers need a safe way to communicate and interact with either other human or machines. For safety reason, traditional visual and tactile man-machine interfaces, such as displays, buttons and knobs are not satisfactory but speech, as the most convenient and natural way of communicate, is an appropriate and complementary solution which can reduce distractions. Hence, Automatic Speech Recognition (ASR) provides safety and convenience, and it is possible to follow the philosophy "Eyes on the road and hands on the steering wheel", which should drive every in-vehicle system design. The problem of robust ASR in car environments has attracted much attention in the recent years and a new market demands for systems which allow the driver to control non critical devices or tasks like phone dialing, RDs-tuner, air conditioner, satellite navigation systems, remote information Web browsing... For this purpose, hands-free interaction in challenging acoustic environments still needs to be improved with respect to several kinds of variabilities.

ASR system performance can be degraded by two important factors: the speaker variability and the acoustic environment. It can be assumed that the speaker variability produces, mainly, a rotation of the feature vectors, while the acoustic environment generates mainly a translation of the feature vectors. In this work we propose a combination of two techniques: augMented stAte space acousTic modEl (MATE), which compensates, by adapting the acoustic models, the rotation effect, and Multi-Environment Model based LInear Normalization (MEMLIN), which is a feature vector normalization technique that obtains important improvements compensating the translations.

The speaker variability problem has been addressed by many authors, specially in the sense of compensation of vocal tract shape by means of the well known Vocal Length Tract Normalization (VLTN) [1] and Maximum Likelihood Linear Regression (MLLR) methods [2]. Those methods still have limitations in order to adapt the acoustic models to the speaker. Usually a great amount of speaker data and exact transcriptions or previous utterances and ASR transcriptions are needed. In this research line, MATE [3] consists of an expansion of the VTLN methods that provides the spectral warping to be locally optimized and simultaneously to the decoding of the state sequence. MATE obtains expanded acoustic models from reference ones using linear transformations and it was proved to be effective in noise free or moderately noisy speech conditions [3], [4]. However the accuracy of a speech recognition system based on MATE with noisy signal rapidly degrades. To compensate this limitation, robustness techniques can be used.

MEMLIN [5] is an effective empirical feature vector normalization technique which compensates the effects of dynamic and adverse acoustic environments. MEMLIN is based on Minimum Mean Square Error (MMSE) estimator, and models clean and noisy spaces assuming Gaussian Mixture Models (GMMs). A bias vector transformation for each pair of Gaussians from the clean and the noisy spaces is defined to compensate the mismatch between clean and noisy feature vectors.

This paper is organized as follows: In Section II, a novel point of view of MATE analysis is explained. In Section III an overview of MEMLIN is detailed. The MATE-MEMLIN algorithm is presented in Section IV. The normalized space acoustic models are explained in Section V. The results with Spanish SpeechDat Car database [6] are included in Section VI, and finally, the conclusions are presented in Section VII.

II. MATE

The main motivation in MATE is to find an acoustic model able to capture speaker variability. MATE provides a mechanism based on the VTLN spectral warping procedure to frame by frame optimized. The acoustic model captures local frequency deformations of the spectrum envelope, which are known to have their origin in the vocal tract

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

A. Miguel, L. Buera, E. Lleida, A. Ortega and O. Saz are with Communication Technologies Group (GTC), Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain. {amiguel,lbuera,lleida,ortega,oskarsaz}@unizar.es

and articulatory instant shapes. A more complex and flexible speech production scheme can be assumed, in which local elastic deformations of the speech can be captured or generated by the model by means of linear transformations, i.e. rotations. Inertia and memory constraints are imposed on the dynamics of the local transformations, then the plausible transformation sequence is assumed to follow an HMM process.

In [7], it was shown that the spectral warping performed by VTLN methods is equivalent to a linear projection of the cepstral feature space. So, for a discrete set of N possible warping factors, α_n, the equivalent MATE transformation matrices $\{\mathbf{A}_n\}_{n=1}^{N}$ can be obtained as

$$\mathbf{V}^{\alpha_n} = \mathbf{A}_n \mathbf{W}, \tag{1}$$

where $n \in [1, N]$ is the index of the warping factor, $\mathbf{W}$ is a matrix which is composed by the source space data, and the $\mathbf{V}^{\alpha_n}$ matrix includes the target space data, which is obtained from the source space data normalized with VTLN using the corresponding α_n warping factor [1].

MATE [3] expands each state of the source space acoustic models (HMM). So, an original state q ($q \in [1, Q]$) will be expanded N times into states (q, n). Thus MATE provides observation generation probability density functions (pdfs) in the states that depend on the discrete set of transformation matrices, $\mathbf{A}_n$, embedding the warping in the acoustic model as a general transformation. Assuming that a component in the pdf mixture of the original state q follows a normal distribution: $\mathcal{N}(\mathbf{x}_t; \mu_q, \mathbf{\Sigma}_q)$, the corresponding expanded state component is assumed to follow the distribution $\mathcal{N}(\mathbf{x}_t; \mathbf{A}_n\mu_q, \mathbf{A}_n\mathbf{\Sigma}_q\mathbf{A}_n^t)$. So, the pdf for the expanded state (q, n), $f(\mathbf{x}_t|n, q)$, is a GMM of the defined expanded components where the a priori component weights remain unaltered.

The expanded acoustic model, from the perspective of a feature vector generator, can be seen as a more flexible speech production process because it can generate sequences of warped cepstrum vectors. To complete the parameter set of the expanded model, the expanded state transition probabilities $\mathbf{\Pi}$, are

$$\mathbf{\Pi} = \{\pi_{q',n',q,n}\}_{q'=1,n'=1,q=1,n=1}^{Q,N,Q,N}, \tag{2}$$

being $\pi_{q',n',q,n}$ the transition probability from state (q', n') to (q, n), which is obtained as [10].

The search algorithm for decoding unlabeled sequences under this framework can be performed by computing recursively the score state variable, $\phi_{q,n}(t)$, for the state (q, n), the index of the warping factor n and the frame time index t.

$$\phi_{q,n}\left(t\right) = \max_{n',q'} \left\{\phi_{q',n'}\left(t-1\right) \cdot \pi_{q',n',q,n}\right\} \cdot f\left(\mathbf{x}_t|n, q\right). \tag{3}$$

This recursive expression is very similar to the one considered in [3], being the main difference how the warping is applied, since now is the expanded acoustic model which tries to generate or evaluate the warped data instead of normalizing the data to fit the source space acoustic model as in [10]. Besides, in the new framework the covariance is normalized in the expanded model description, including the Jacobian normalization in the model [7].

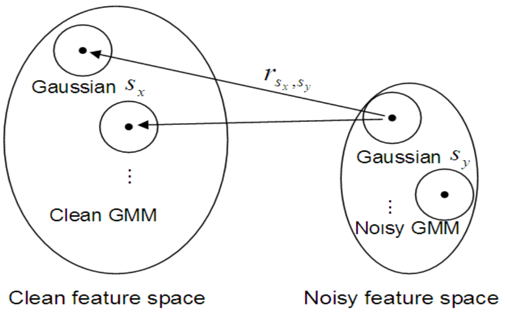

Fig. 1. Scheme of MEMLIN approximations for one basic environment, where s_x and s_y are the index of clean and noisy space Gaussians and $\mathbf{r}_{s_x,s_y}$ is the bias vector transformation associated to the pair of Gaussians s_x and s_y.

A. *MSE Transformation matrix estimation*

The rotation MATE matrices $\mathbf{A}_n$, as a general linear transformation for the feature vectors, provide a great degree of freedom for the MATE expanded models, including the rotation transformations in the mean vectors and covariance matrices.

In order to estimate the rotation matrices $\mathbf{A}_n$, a linear transformation as (1) is defined and the multidimensional regression Minimum Square Error (MSE) criterion is used in a previous process with training data. So, a residual error is defined as the squared sum of differences between the target data, $\mathbf{V}^{\alpha_n}$ $(D \times L)$ (D is the dimension of the feature vectors and L is the number of feature vectors) and the projected ones, $\mathbf{A}_n\mathbf{W}$, where $\mathbf{W}$ matrix includes the source data $(D \times L)$. Taking derivatives with respect to $\mathbf{A}_n$ and equating them to zero, we obtain the following expression for the estimation of $\mathbf{A}_n$

$$\mathbf{A}_n = (\mathbf{W}\mathbf{W}^t)^{-1}\mathbf{W}(\mathbf{V}^{\alpha_n})^t. \tag{4}$$

III. MEMLIN OVERVIEW

MEMLIN is an empirical feature vector normalization technique based on MMSE estimator. It assumes three approximations: the clean feature space is modelled as a mixture of Gaussians (GMM), the noisy one is split into several basic acoustic environments and each one of them is modelled as a GMM. 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. These assumptions can be shown, in a schematic way, in Fig. 1 for one basic environment, where clean and noisy spaces are modelled by GMMs and the corresponding bias vector transformation between a clean space GMM component (s_x) and a noisy space GMM component (s_y) is depicted as r_{s_x,s_y}.

A. MEMLIN approximations

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

$$f(\mathbf{x}_t) = \sum_{s_x} f(\mathbf{x}_t|s_x)p(s_x), \tag{5}$$

$$f(\mathbf{x}_t|s_x) = \mathcal{N}(\mathbf{x}_t; \mu_{s_x}, \mathbf{\Sigma}_{s_x}), \tag{6}$$

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.

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

$$f_e(\mathbf{y}_t) = \sum_{s_y^e} f(\mathbf{y}_t|s_y^e)p(s_y^e), \tag{7}$$

$$f(\mathbf{y}_t|s_y^e) = \mathcal{N}(\mathbf{y}_t; \mu_{s_y^e}, \mathbf{\Sigma}_{s_y^e}), \tag{8}$$

where s_y^e denotes the corresponding Gaussian of the noisy model for the e basic environment, $\mu_{s_y^e}$, $\mathbf{\Sigma}_{s_y^e}$ 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 vectors, 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.

B. MEMLIN enhancement

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

$$\hat{\mathbf{x}}_t = \mathbf{y}_t - \sum_e \sum_{s_y^e} \sum_{s_x} \mathbf{r}_{s_x,s_y^e} p(e|\mathbf{y}_t) p(s_y^e|\mathbf{y}_t, e) p(s_x|\mathbf{y}_t, e, s_y^e), \tag{9}$$

where $p(e|\mathbf{y}_t)$ is the a posteriori probability of the basic environment e; $p(s_y^e|\mathbf{y}_t, e)$ is the a posteriori probability of the noisy model Gaussian s_y^e, given the noisy feature vector $\mathbf{y}_t$ and the basic environment e. Those two terms are computed on-line for each frame applying (7) and (8), as described in [5]. 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 noisy feature vector $\mathbf{y}_t$, the basic environment e, and the noisy model Gaussian s_y^e. That term, along with the bias vector transformation $\mathbf{r}_{s_x,s_y^e}$, is estimated in a previous training phase using stereo data [5]. If stereo data is not available, a “blind” version of the training phase can be applied [9].

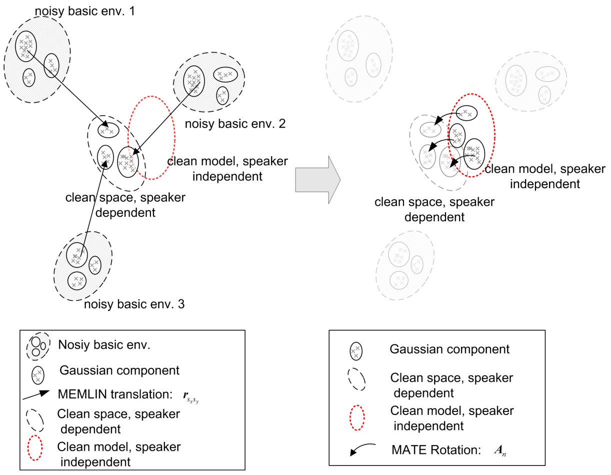
noisy basic env. 1
noisy basic env. 2
clean model, speaker independent
clean space, speaker dependent
noisy basic env. 3
clean model, speaker independent
clean space, speaker dependent
Nosiy basic env.
Gaussian component
MEMLIN translation: $r_{s_x s_y}$
Clean space, speaker dependent
Clean model, speaker independent
Gaussian component
Clean space, speaker dependent
Clean model, speaker independent
MATE Rotation: A_n

Fig. 2. Scheme of MATE-MEMLIN performance.

IV. MATE-MEMLIN

In order to provide robustness to MATE in adverse acoustic conditions, MATE-MEMLIN combines MATE with MEMLIN. So, the MEMLIN normalize feature vectors are decoded using MATE-MEMLIN expanded acoustic models. Since the rotation matrices can be estimated in a data driven way and MEMLIN maps the different basic environment data towards only one space, the corresponding matrices for MATE-MEMLIN can be adapted to the specific problem without considering any environment dependence (see Fig. 2). In the left part of the Fig. 2, the MEMLIN normalization is depicted and the speaker noisy data are mapped to a clean speaker dependent space. The MATE effect is presented on the right part of the figure, where the clean independent speaker acoustic models are adapted to the optimal ($\mathbf{A}_n$) speaker dependent space. This acoustic model election ($\mathbf{A}_n$) is obtained frame by frame in the search algorithm by maximum likelihood (3).

As it has been described in Section II, the MATE rotation matrices obtained with the MSE linear regression criterion need matched source and target data. In this work it is assumed available stereo training data for each basic environment e, $(\mathbf{X}^e, \mathbf{Y}^e) = \{(\mathbf{x}_1^e, \mathbf{y}_1^e), ...(\mathbf{x}_t^e, \mathbf{y}_t^e)..., (\mathbf{x}_T^e, \mathbf{y}_T^e)\}$, where $\mathbf{X}^e$ represents the clean feature vectors for the basic environment e, and $\mathbf{Y}^e$ the corresponding noisy ones (these stereo data are also needed to obtain the bias vector transformations and the cross-probability models for MEMLIN [5]).

In the MATE-MEMLIN rotation matrix estimation, the source space data are $\mathbf{X}$, which is the concatenation of the clean training data for all the basic environments. On the other hand, the target space data are obtained with noisy data $\mathbf{Y}$ for all the basic environments, including rotation and translation compensations as following

- Noisy warped cepstrum feature vectors ($\mathbf{Y}^{\mathbf{e},\alpha_\mathbf{n}}$) are obtained applying the VTLN rotation [1] to the noisy training data for all the basic environments e, $\mathbf{Y}^\mathbf{e}$.
- MEMLIN compensation algorithm is applied over

TABLE I

WER BASELINE RESULTS, IN %, FROM THE DIFFERENT BASIC ENVIRONMENTS (E1,..., E7), WHERE MWER IS THE MEAN WER.

Train	Test	E1	E2	E3	E4	E5	E6	E7	MWER (%)
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

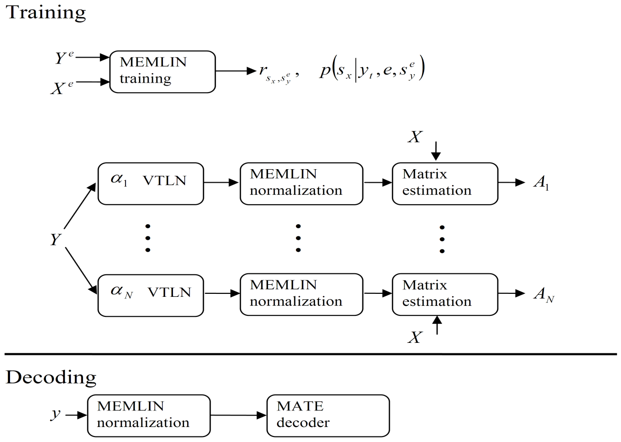

Fig. 3. Scheme of MATE-MEMLIN phases.

$\mathbf{Y^{e,\alpha_n}}$, obtaining the normalized data $\hat{\mathbf{X}}^{e,\alpha_n}$ (9). The target space data $\hat{\mathbf{X}}^{\alpha_n}$ are composed by the concatenation of the normalized data for all basic environments.

Thus, finally, the rotation matrix estimations, which are defined as the following linear projection $\hat{\mathbf{X}}^{\alpha_n} = \mathbf{A}_n\mathbf{X}$, are (see Section II)

$$\mathbf{A}_n = (\mathbf{X}\mathbf{X}^t)^{-1}\mathbf{X}(\hat{\mathbf{X}}^{\alpha_n})^t. \qquad (10)$$

In decoding, the noisy feature vectors are normalized with MEMLIN algorithm and the normalized data are recognized using the MATE-MEMLIN expanded acoustic models with the transformation matrices, $\mathbf{A}_n$ (10). A graphical representation for the rotation matrix estimation method and the recognition process for MATE-MEMLIN can be observed in Fig. 3.

Note that the resulting expanded MATE-MEMLIN acoustic models are able to locally rotate the frequency axis in the standard VTLN way, including at the same time the information of how the MEMLIN normalized feature vectors are distributed in the feature space.

V. NORMALIZED SPACE ACOUSTIC MODELS

Feature vector normalization techniques try to map the noisy feature vectors to the clean space. However this mapping is not perfect and a new normalized space is created, which is different to the clean one. So, a further improvement can be obtained adapting the clean acoustic models towards the normalized space. For this purpose, the noisy training data are normalized in the same way as testing data and the original clean acoustic models are adapted with those data towards the new normalized space. If there are enough data, Maximum Likelihood (ML) algorithm can be used, but a model adaptation method should be applied otherwise (Maximum A Posteriori, MAP [11], MLLR [2]...). In this work, once the MEMLIN normalized space acoustic models are obtained, the normalized testing data can be recognized directly with them, or with new expanded MATE-MEMLIN acoustic models. In this case the normalized acoustic models are expanded with the corresponding MATE transformation matrices. In this work, the two options are compared.

VI. RESULTS

To compare the performance of the MATE-MEMLIN technique in a real, dynamic, and complex car environment, a set of experiments were carried out using the Spanish SpeechDat Car database [6]. 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).

Two channels of the database, which were recorded simultaneously (stereo data), were used: a clean signal from a CLose talK channel (CLK), which was recorded with a Shure SM-10A microphone, and a noisy signal from a Hands-Free channel (HF), which was recorded using a Peiker ME15/V520-1 microphone located on the ceiling in front of the driver. HF signals were used in recognition tasks.

The SNRs (mean ± standard deviation) of the HF channel range from 14.05±3.89 dB in the E1 basic environment to 5.65±4.35 dB in the high speed and good road conditions (E6 and E7 basic environments combined).

The recognition task is isolated and continuous digits recognition (a typical hands-free phone task). As feature set, the standard ETSI front-end features plus the energy and the corresponding delta and delta delta coefficients were used in all the experiments [12]. Cepstral mean normalization is applied to testing and training data in all cases. On the other hand, in this work, the VTLN, MEMLIN and SPLICE with environmental model selection [8] (SPLICE MS) algorithms were applied to the 12 MFCCs and energy, whereas the derivatives were computed over the normalized static coefficients. The acoustic models were composed of 16 state HMM for each digit, a 3 state begin-end silence HMM and an 1 state inter-word silence HMM. In all cases, each pdf state is composed by a mixture of three Gaussians.

A. Baseline results

The Word Error Rate (WER) baseline results for each basic environment are presented in Table I, where MWER is the Mean WER, which is computed proportionally to the number of words on each basic environment. "Train" column refers to the signals used to obtain the corresponding acoustic models: if they are trained with all clean training utterances, the column is marked CLK, and if the column is marked HF, the acoustic models are trained with all noisy training utterances. "Test" column indicates the signals which are used for recognition: clean, CLK, or noisy, HF.

Table I shows the effect of real car noise, which produces a significant increase in WER in all of the basic environments, (Train CLK, Test HF), concerning the rates for clean conditions, (Train CLK, Test CLK). With matched conditions: when acoustic models are retrained using all basic environments, (Train HF, Test HF) MWER decreases significantly.

B. MATE, MEMLIN and MATE-MEMLIN results

Table II shows the MWER when only MATE is applied. The expanded MATE models are obtained over the ones trained with "Train" column signals. To compute the transformation matrices, the source space is the clean one, and the target space is obtained with the clean data normalized with VTLN using 5 warping factors (0.8, 0.9, 1.0, 1.1 and 1.2 [1]).

TABLE II
MEAN WER (MWER) IN % FROM MATE TECHNIQUE.

Train	Test	MWER (%)
MATE CLK	CLK	0.76
MATE CLK	HF	29.28
MATE HF	HF	7.30

It can be verified in Table II the improvement that expanded MATE models obtain when they are applied over the clean signals (0.76% of MWER) concerning the result obtained when clean feature vectors are recognized with clean acoustic models (0.91% of MWER). This result is better than if VTLN [1] is applied over clean signal (0.81% of MWER). Furthermore, the basic VTLN technique is not on-line as MATE. However, MATE is not very effective in noisy conditions due to the high noise sensibility of the method. This the reason of using MEMLIN combined with MATE, because they are complementary and the bad performance of MATE in noisy conditions can be compensated with MEMLIN.

To compensate the effects of the noise in recognition, MEMLIN and MATE-MEMLIN are proposed. Fig. 4 shows the mean improvement in WER (MIMP) in % for MEMLIN and MATE-MEMLIN. Also the results obtained with SPLICE MS are included to compare. Given a Mean WER, the corresponding mean improvement in WER is computed as

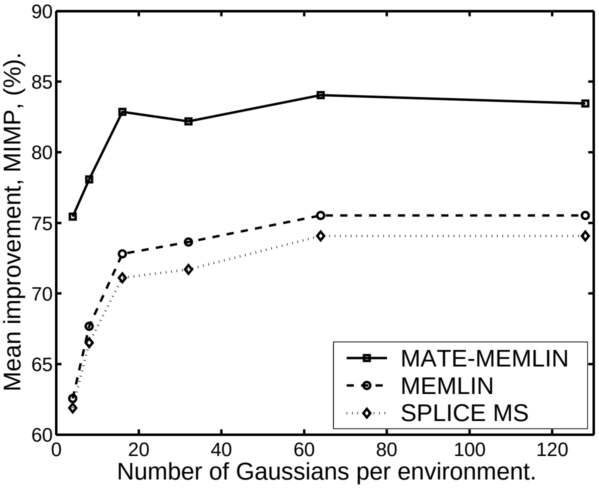

Fig. 4. Mean improvement in WER (MIMP) in % for MATE-MEMLIN, MEMLIN and SPLICE with environmental model selection (SPLICE MS) when different number of Gaussians per basic environment are considered.

$$MIMP = \frac{100(MWER - MWER_{CLK-HF})}{MWER_{CLK-CLK} - MWER_{CLK-HF}}, \tag{11}$$

where $MWER_{CLK-CLK}$ is the mean WER obtained with clean conditions (0.91 in this case), and $MWER_{CLK-HF}$ is the baseline (21.48). So, A 100% MIMP would be achieved when MWER equals the one obtained under clean conditions. In order to compare all the methods, the MIMP has been depicted with respect to the number of Gaussians per basic environment because it gives an idea of the computing cost. It can be observed the improvement of MATE-MEMLIN with respect to MEMLIN: from 62.57% to 75.44% (from 8.61% to 5.96% of MWER) with only 4 Gaussians per basic environment and from 75.53% to 83.45% (from 5.95% to 4.32% of MWER) with 128 Gaussians, showing the importance of the use of the expanded MATE-MEMLIN acoustic models with after MEMLIN normalization. On the other hand, note the important improvement of MATE-MEMLIN with respect to SPLICE MS. It is also important to observe that the dependence of the results concerning the number of Gaussians per basic environment has been reduced when MATE-MEMLIN is applied; so competitive results can be obtained with a few number of Gaussians per basic environment. The best results of MWER for SPLICE MS, MEMLIN and MATE-MEMLIN (all of them obtained with 64 Gaussians per basic environment) are included in Table III

C. Results with normalized space acoustic models

Table IV shows the corresponing matching condition results (MWER and MIMP) when normalized acoustic models are used. Clean and noisy condition results (Train CLK, Test CLK and Train HF, Test HF, respectively) are included again

TABLE III

BEST MEAN WER (MWER) IN % FROM SPLICE WITH ENVIRONMENTAL MODEL SELECTION (SPLICE MS), MEMLIN AND MATE-MEMLIN TECHNIQUES (ALL OF THEM OBTAINED WITH 64 GAUSSIANS PER BASIC ENVIRONMENT).

Train	Test	MWER (%)
CLK	HF SPLICE MS	6.25
CLK	HF MEMLIN	5.95
CLK	HF MATE-MEMLIN	4.19

to compare. In “HF MEM” the normalized with MEMLIN noisy training data are used to retrain the new acoustic models with the ML algorithm. The results for “HF MAT-MEM” are obtained with the expanded MATE-MEMLIN acoustic models estimated from the ones retrained with the MEMLIN normalized noisy training data and the ML algorithm. In both cases MEMLIN is applied with 128 Gaussians per basic environment. The recognition rates using less number of Gaussians for MEMLIN are similar, for instance, if 4 Gaussians per basic environment are used in MEMLIN, the MIMP with the new retrained acoustic models is 94.63%, and if the corresponding expanded acoustic models are estimated by MATE-MEMLIN, the MIMP reaches 95.06%. Clearly there are significant improvements when normalized space acoustic models are used, even when noisy acoustic models are applied. Furthermore, in this case the number of Gaussians used for MEMLIN does not affect significatively to the performance.

TABLE IV

MEAN WER (MWER) AND MEAN IMPROVEMENT IN WER (MIMP) IN % FROM MEMLIN (MEM) AND MATE-MEMLIN (MAT-MEM) WITH 128 GAUSSIANS PER BASIC ENVIRONMENT AND ML-ADAPTED ACOUSTIC MODELS TO THE NORMALIZED SPACE.

Train	Test	MWER (%)	MIMP (%)
CLK	CLK	0.91	–
HF	HF	4.63	81.93
HF MEM	HF MEM	1.72	96.08
HF MAT-MEM	HF MAT-MEM	1.68	96.25

VII. CONCLUSIONS

A. Conclusions

In this paper we have presented the MATE-MEMLIN, which is a combination between a novel point of view of MATE (acoustic model adaptation technique) and MEMLIN (feature vector normalization algorithm), in order to compensate the speaker variability and the car environment effects. MATE proposes new expanded acoustic models from original models to normalize the vocal track length. Some results with Spanish SpeechDat Car database show the effective behaviour of the technique with clean signals, 0.76% of MWER, (better than VTLN, which reaches 0.81% of MWER), but, with noisy data the system rapidly degrades. To compensate this mismatch, the feature vector normalization technique MEMLIN is selected to normalize the noisy data, defining MATE-MEMLIN. So, in MATE-MEMLIN, the MEMLIN normalize feature vectors are decoded using MATE-MEMLIN expanded acoustic models. MATE-MEMLIN obtains an improvement in WER of 83.45% with 128 Gaussians per basic environment, whereas MEMLIN in the same conditions reaches 75.53%. If expanded normalized space acoustic models are used in recognition, the mean improvement is 96.25% with 128 Gaussians per basic environment.

REFERENCES

[1] L. Lee and R. Rose, ”A Frequency Warping Approach to Speaker Normalization”, *in IEEE Transactions on Speech and Audio Processing*, 1998, volume 1, number 6, pp. 49-60.

[2] C.J. Leggetter and P.C. Woodland, ”Maximum Likelihood Linear Regression for Speaker Adaptation of continuous density hidden Markov models”, *in Computer Speech and Language*, 1995, volume 9, pp. 171-185.

[3] A. Miguel and E. Lleida and R. Rose and L. Buera and A. Ortega, ”Augmented State Space Acoustic Decoding for Modeling Local Variability in Speech”, *in Proceedings of Eurospeech*, 2005, pp. 3009-3012, Lisbon, Portugal.

[4] R. Rose and A. Keyvani and A. Miguel, ”On the Interaction Between Speaker Normalization, Environment Compensation, and Discriminant Feature Space Transformations”, *in Proceedings of ICASSP*, 2006, Toulouse, France.

[5] L. Buera and E. Lleida and A. Miguel and A. Ortega, ”Multi-Environment models based linear normalization for speech recognition in car conditions”, *in Proceedings of ICASSP*, 2004, Motreal, Canada.

[6] A. Moreno and B. Lindberg and C. Draxler and G. Richard and K. Choukri and S. Euler and J. Allen, ”SPEECHDAT-CAR. A large speech database for automotive environments”, *in Proceedings of LREC*, 2000, vol. 2, pp. 895-900, Athens, Greece.

[7] M. Pitz and H. Ney, ”Vocal Tract Normalization Equals Linear Transformation in Cepstral Space”, *in IEEE Transactions on Speech and Audio Processing*, 2005, vol. 13, number 5, pp. 930-944.

[8] J. Droppo and L. Deng and A. Acero, ”Evaluation of the SPLICE algorithm on the AURORA2 database”, *in Proceedings of EUROSPEECH*, 2001, vol. 1, pp. 217-220, Aalborg, Denmark.

[9] L. Buera and E. Lleida and A. Miguel and A. Ortega, ”Recent advances in PD-MEMLIN for speech recognition in car conditions”, *in Proceedings of ASRU*, 2005, pp. 180-185. San Juan, Puerto Rico.

[10] A. Miguel and E. Lleida and A. Juan and L. Buera and A. Ortega and O. Saz, ”Local Transformation Models for Speech Recognition”, *in Proceedings of ICSLP*, 2006, Pittsburgh, USA.

[11] J.-L. Gauvain and C.-H. Lee, ”Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains”, *IEEE Transactions on Speech and Audio Processing*, april 1994, vol. 2, pp 291-298.

[12] ETSI, ”Speech Processing Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms”, *ETSI ES 201 108 version 1.1.2*, April 2000.