INTERSPEECH 2009 BRIGHTON

Unsupervised Training Scheme with Non-Stereo Data for Empirical Feature Vector Compensation

L.Buera[1] *, *A.Miguel*[1], *A.Ortega*[1], *E.Lleida*[1], *R.M.Stern*[2]

[1]Communication Technologies Group (GTC)
I3A, University of Zaragoza, Spain
[2]Dep. Electrical and Computer Engineering
Carnegie Mellon University, USA
{lbuera,amiguel,ortega,lleida}@unizar.es, rms@cs.cmu.edu

Abstract

In this paper, a novel training scheme based on unsupervised and non-stereo data is presented for Multi-Environment Model-based LInear Normalization (MEMLIN) and MEMLIN with cross-probability model based on GMMs (MEMLIN-CPM). Both are data-driven feature vector normalization techniques which have been proved very effective in dynamic noisy acoustic environments. However, this kind of techniques usually requires stereo data in a previous training phase, which could be an important limitation in real situations. To compensate this drawback, we present an approach based on ML criterion and Vector Taylor Series (VTS). Experiments have been carried out with Spanish SpeechDat Car, reaching consistent improvements:48.7% and 61.9% when the novel training process is applied over MEMLIN and MEMLIN-CPM, respectively.

Index Terms: unsupervised non-stereo training data, feature vector normalization, MEMLIN.

1. Introduction

Automatic Speech Recognition (ASR) systems can achieve high performance in controlled conditions. However, when training and recognition acoustic conditions differ, the accuracy of the systems rapidly degrades. To compensate for the effects which cause the mismatch between training and recognition spaces, robustness techniques have been developed [1].

This work is focussed on empirical data-driven feature vector normalization methods, which provide a dynamic approach that requires less data and computing time than acoustic model adaptation methods do. However, most of these techniques demand stereo data in a previous training process, which may be an important limitation in real situations. To compensate this drawback, some solutions have been explored. Thus, Maximum Likelihood (ML) framework was applied for multivariate Gaussian-based cepstral normalization (RATZ) [2]. Also discriminative training was used for Stereo-based Piecewise Linear compensation for Environments (SPLICE) [3]. However, note that this last solution is not unsupervised since the transcription of the training data is required, so that an active enrollment by the speakers is needed during the training process.

In this work, an unsupervised non-stereo data training approach based on ML criterion is presented for Multi-Environment Model-based LInear Normalization (MEMLIN) [4]. In a similar way as [2], Expectation Maximization (EM) technique [5] is used to provide the ML solution for all the parameters that have to be estimated. However, instead of using just the MEMLIN-based signal model, which relates the noisy and clean feature vectors, Vector Taylor Series (VTS) [2] [6] is also applied. To compare the performance of the proposed training method, several experiments with Spanish SpeechDat Car database [7] were carried out, reaching 61.9% of average improvement.

This paper is organized as follows: In Section 2, a brief overview of MEMLIN is included. The proposed unsupervised non-stereo data training scheme is presented in Section 3. In Section 4, the comparative results with Spanish SpeechDat Car database are detailed. Finally, the conclusions and future work are discussed in Section 5.

This work has been supported by the national project TIN08-06856-C05-04. Recently Luis Buera has joined Speech Technology Group at Toshiba Research Europe Ltd.

2. MEMLIN overview

Multi-Environment Model-based LInear Normalization (MEMLIN) is an empirical feature vector normalization technique based on a general MMSE framework where clean and noisy spaces are modelled each with a GMM.

2.1. MEMLIN approximations

• Clean feature vectors, $\mathbf{x}$, are modelled with a GMM of N_x components.

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

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

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.

• Noisy space is split into several basic environments e which represent different acoustic conditions. Furthermore, the corresponding feature vectors $\mathbf{y}$ are modeled as a GMM of N_y components for each basic environment

$$p(\mathbf{y}|e) = \sum_{s_y^e=1}^{N_y} p(s_y^e|e)p(\mathbf{y}|s_y^e, e), \tag{3}$$

$$p(\mathbf{y}|s_y^e, e) = \mathcal{N}(\mathbf{y}; \mu_{s_y^e}, \mathbf{\Sigma}_{s_y^e}), \tag{4}$$

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 $p(s_y^e|e)$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated to s_y^e.

• Finally, the relationship between $\mathbf{x}$ and $\mathbf{y}$ (signal model) is considered linear within each pair of Gaussians s_x and s_y^e: $p(\mathbf{x}|\mathbf{y}, s_x, s_y^e, e) = \mathcal{N}(\mathbf{x}; \mathbf{y} - \mathbf{r}_{s_x,s_y^e}, \mathbf{\Sigma}_{s_x})$, where $\mathbf{r}_{s_x,s_y^e}$ is the bias vector transformation between noisy and clean feature vectors associated to each pair of Gaussians (s_x and s_y^e).

Although MEMLIN proposes a linear signal model based just on a bias vector, different approximations can be considered, such as first order polynomial or even non linear estimates [4].

2.2. MEMLIN enhancement

In order to estimate the clean feature vector for each time index t ($\hat{\mathbf{x}}_t$), the MMSE estimator is applied combining the three approximations above

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

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 noisy feature vector $\mathbf{y}_t$ and the basic environment e. Those two terms are computed for each testing feature vector combining (3) and (4) in the recognition phase [4]. 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 $\mathbf{r}_{s_x,s_y^e}$ has to be estimated in a previous unsupervised training process.

3. Non-stereo based ML training process

Given a noisy training corpus for each basic environment, $\mathbf{Y}_e = \{\mathbf{y}_1^e; ...; \mathbf{y}_{t_e}^e; \ ...; \mathbf{y}_{T_e}^e\}$, with $t_e = 1, ..., T_e$, the bias vector transformation $\mathbf{r}_{s_x,s_y^e}$ and the cross-probability model $p(s_x|\mathbf{y}_t, e, s_y^e)$ are estimated with ML criterion, applying EM algorithm [5]. Hence, the cross-probability model is simplified avoiding the time dependence given by the noisy feature vector $\mathbf{y}_t$ (static cross-probability model: $p(s_x|\mathbf{y}_t, e, s_y^e) \simeq p(s_x|e, s_y^e)$). Thus, assuming the approximations of subsection 2.1, the expected value of the complete-data log-likelihood, given the observed data $\mathbf{Y}_e$ and the current parameter estimates $\theta_e^{(k)}$ can be obtained as (6). Observe that $\theta_e^{(k)}$ includes the variables $p(s_x|e, s_y^e)^{(k)}$ and $\mathbf{r}_{s_x,s_y^e}^{(k)}$, (k) represents the index of iteration and $z_{t_e,s_x,s_y^e}^{(k)}$ is the expected value of the Gaussians s_x and s_y^e, given the noisy feature vector $\mathbf{y}_{t_e}$ and $\theta_e^{(k)}$: $z_{t_e,s_x,s_y^e}^{(k)} = E[s_x, s_y^e|\mathbf{y}_{t_e}, \theta_e^{(k)}]$. Assuming the constraint $\sum_{s_x} p(s_x|e, s_y^e) = 1$, and deriving with respect to the corresponding variables, the ML solutions for the bias vector transformation and the static cross-probability model can be computed as

$$p(s_x|e, s_y^e)^{(k+1)} = \frac{\sum_{t_e} z_{t_e,s_x,s_y^e}^{(k)}}{\sum_{t_e} \sum_{s_x} z_{t_e,s_x,s_y^e}^{(k)}}. \quad (7)$$

$$\mathbf{r}_{s_x,s_y^e}^{(k+1)} = \frac{\sum_{t_e} z_{t_e,s_x,s_y^e}^{(k)} (\mathbf{y}_{t_e} - \mu_{s_x})}{\sum_{t_e} z_{t_e,s_x,s_y^e}^{(k)}}. \quad (8)$$

In order to estimate $z_{t_e,s_x,s_y^e}^{(k)}$, independence between Gaussians is assumed: $z_{t_e,s_x,s_y^e}^{(k)} \simeq E[s_y^e|\mathbf{y}_{t_e}, \theta_e^{(k)}] \times E[s_x|\mathbf{y}_{t_e}, \theta_e^{(k)}]$. Thus, $E[s_y^e|\mathbf{y}_{t_e}, \theta_e^{(k)}]$ is computed using (3) and (4) as $p(s_y^e|\mathbf{y}_{t_e})$. Following a similar approach to [2], $E[s_x|\mathbf{y}_{t_e}, \theta_e^{(k)}]$ could be evaluated using (3), (4) and $\mathcal{N}(\mathbf{y}_{t_e}; \mu_{s_x} + \mathbf{r}_{s_x,s_y^e}^{(k)}, \mathbf{\Sigma}_{s_x})$ for all noisy model Gaussians. Observe that the last expression can be seen as the likelihood of the noisy feature vector when s_x has been adapted to the noisy space by using the MEMLIN signal model, so that the clean model mean is shifted with $\mathbf{r}_{s_x,s_y^e}^{(k)}$ for a given s_y^e.

However, preliminary results showed us that better performance could be obtained by adapting the clean space GMM to the corresponding noisy environment using VTS [2] [6] instead of the MEMLIN signal model. Thus, for a given training utterance of e basic environment, each clean model Gaussian s_x is transformed into a noisy model Gaussian $s_{x,VTS}^e$ using VTS approach. So $E[s_x|\mathbf{y}_{t_e}, \theta_e^{(k)}] \simeq p(s_{x,VTS}^e|\mathbf{y}_{t_e})$. Although sometimes the assumed degradation function for VTS (convolutional distortion and additive noise) could be unsuitable in real situations, the technique provides a reasonable solution to estimate $E[s_x|\mathbf{y}_{t_e}, \theta_e^{(k)}]$ due to the relation one to one between the clean model and the the adapted model Gaussians. Furthermore, observe that this approach provides a non iterative solution.

Note that if stereo data are available, $(\mathbf{Y}_e, \mathbf{X}_e) = \{(\mathbf{y}_1^e, \mathbf{x}_1^e); ...; (\mathbf{y}_{t_e}^e, \mathbf{x}_{t_e}^e); \ ...; (\mathbf{y}_{T_e}^e, \mathbf{x}_{T_e}^e)\}$, non iterative solutions are obtained [4]: $z_{t_e,s_x,s_y^e}^{(k)}$ can be estimated as $p(s_y^e|\mathbf{y}_{t_e}, e)\ p(s_x|\mathbf{x}_{t_e})$ and μ_{s_x} is replaced by $\mathbf{x}_{t_e}$ in (8).

In a previous work [8], it was shown the big impact of the cross-probability model in the performance of MEMLIN. Also, it was proposed a non static approach based on modelling the noisy feature vectors associated to each pair of Gaussians (s_x and s_y^e) with a new GMM of N_y' components: $p(\mathbf{y}_t|s_x, s_y^e, e, s_y') = \mathcal{N}(\mathbf{y}_t; \mu_{s_x,s_y^e,s_y'}, \mathbf{\Sigma}_{s_x,s_y^e,s_y'})$, where $\mu_{s_x,s_y^e,s_y'}$ and $\mathbf{\Sigma}_{s_x,s_y^e,s_y'}$ are the mean and the diagonal covariance matrix corresponding to the s_y' Gaussian of the cross-probability model associated to s_x and s_y^e. To obtain these two parameters and $p(s_y'|s_x, s_y^e, e)$, which is the a priori probability associated to s_y', a previous training process with stereo data for each basic environment was applied using ML criterion. Hence, each noisy feature vector associated to a basic environment e was labeled with the most probable noisy and clean model Gaussians ($\hat{s}_y^e$ and $\hat{s}_x$). This technique is called MEMLIN with cross-probability model based on GMMs (MEMLIN-CPM). The expansion of this solution for non stereo data training process is straightforward using the labels $\hat{s}_y^e$ and $\hat{s}_{x,VTS}^e$.

Once the cross-probability GMM parameters are estimated, $p(s_x|\mathbf{y}_t, s_y^e, e)$ can be computed in the recognition phase as

$$p(s_x|\mathbf{y}_t, e, s_y^e) = \frac{p(\mathbf{y}_t|s_x, s_y^e, e) p(s_x|s_y^e, e)}{\sum_{s_x} p(\mathbf{y}_t|s_x, s_y^e, e) p(s_x|s_y^e, e)}, \quad (9)$$

$$Q(\theta_e|\theta_e^{(k)}) = \sum_{t_e}\sum_{s_x}\sum_{s_y^e} z_{t_e,s_x,s_y^e}^{(k)} \left[log(\mathcal{N}(\mathbf{y}_{t_e}; \mu_{s_x} + \mathbf{r}_{s_x,s_y^e}, \mathbf{\Sigma}_{s_x})) + log(p(s_y^e)) + log(p(s_x|e, s_y^e)) \right]. \quad (6)$$

Train	Test	E1	E2	E3	E4	E5	E6	E7	AWER (%)
CLK	CLK	0.6	2.4	0.7	0.3	1.0	0.5	0.3	1.0
CLK	HF	2.0	7.5	5.9	8.2	13.5	6.7	29.6	8.5
HF	HF	0.9	4.3	1.7	2.4	3.3	2.4	1.0	2.5

Table 1: *WER baseline results, in %, from the different basic environments (E1,..., E7). AWER is the Average WER.*

$$p(\mathbf{y}_t|s_x, s_y^e, e) = \sum_{s_y^{'}} p(s_y^{'}|s_x, s_y^e, e) p(\mathbf{y}_t|s_x, s_y^e, e, s_y^{'}). \quad (10)$$

Note that the time-independent assumption considered previously has been avoided.

4. Results

To study the performance of the proposed unsupervised non stereo data training scheme, a set of experiments were carried out using the Spanish SpeechDat Car database [7], which is composed of real, dynamic, and complex environments. 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), while 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 recognition task is isolated and continuous digits recognition. As feature set, the standard MFCC front-end features (C0 to C12) are used. Also delta- and delta-delta coefficients are included to complete the 39-dimension feature vectors. On-line cepstral mean normalization is applied to testing and training data. The feature vector normalization techniques are applied over the 13 MFCCs, 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 of a mixture of 3 Gaussians components, except begin-end silence, whose pdf states are composed by 6 Gaussian components.

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 words of each basic environment. The "Train" column refers to the data used to train the corresponding acoustic HMMs: clean training utterances (CLK), or noisy ones (HF), which represents multi-condition training. All acoustic models are trained with ML algorithm. "Test" column indicates which signals are used for decoding: clean (CLK) or noisy (HF). Table 1 shows the effect of real car conditions, which produces a significant increase in WER for all of the basic environments, (compare Train CLK, Test HF with Train CLK, Test CLK). When acoustic models are retrained (ML criterion) using all basic environment signals (Train HF), the AWER decreases considerably to 2.5%.

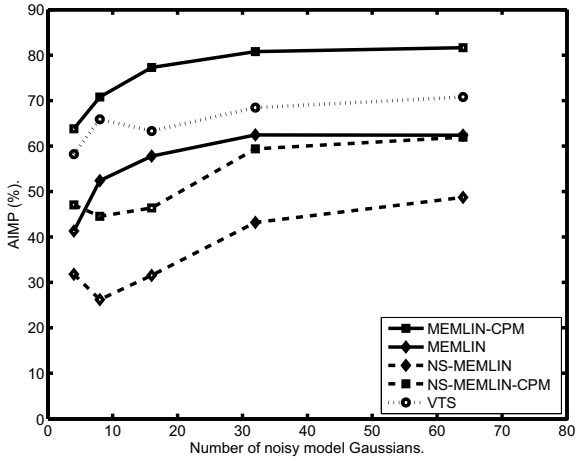

Figure 1: *Mean improvement in WER, MIMP, in % for different normalization techniques.*

In order to study the performance of the proposed training approach, the Average Improvement in WER (AIMP), in %, is defined. Thus, given an AWER, the corresponding AIMP is computed as:

$$AIMP = \frac{100(AWER - AWER_{CLK-HF})}{AWER_{CLK-CLK} - AWER_{CLK-HF}}, \quad (11)$$

where $AWER_{CLK-CLK}$ is the average WER obtained with clean conditions (1.0 in this case), and $AWER_{CLK-HF}$ is the baseline (8.5).

Figure 1 shows the AIMP for MEMLIN and MEMLIN-CPM when the proposed unsupervised non-stereo data training scheme is applied (NS-MEMLIN and NS-MEMLIN-CPM, respectively). Different number of Gaussians per basic environment is used ($N_y = N_x$). Note that N_y can give us a qualitative idea of the computational cost of the presented empirical feature vector adaptation techniques in the decoding phase. Also, results with MEMLIN and MEMLIN-CPM are included. Finally, results reached with Vector Taylor Series for feature vector normalization (VTS) [2] are also depicted for comparison. When cross-probability model based on GMMs is applied, 2 components are used ($N_y^{'}$).

It can be verified in Figure 1 how NS-MEMLIN provides an interesting improvement (48.7% AIMP, 4.7% AWER), although

it is still far away from the performance of MEMLIN. However, the performance of NS-MEMLIN-CPM is quite similar to MEMLIN, reaching 61.9% AIMP, 3.8% AWER. Also, some experiments were carried out with RATZ with non-stereo training data [2] for comparison. The most competitive ones are not as satisfactory as NS-MEMLIN-CPM ones (53.6% AIMP, 4.5% AWER). Finally, the best results are obtained when MEMLIN-CPM is applied (81.7% AIMP, 2.4% AWER), which overcomes clearly VTS (70.8% AIMP, 3.2% AWER), although VTS estimates the degradation model of the signal per testing utterance. Note that all the techniques are unsupervised and even under that condition, better results than supervised multi-condition ML training can be achieved (Train HF, Test HF).

Observe that the techniques included in this section have very different requirements, so that it is difficult to compare them. So, MEMLIN and MEMLIN-CPM, which are empirical approaches and obtain very competitive results, require stereo training data, which sometimes are not available. Also, these techniques are sensitive to the mismatch between training and recognition conditions. On the other hand, VTS, whose performance is also quite satisfactory, needs to estimate the degradation model parameters for each testing utterance, so that the computation cost in decoding process is increased. Also, VTS may have some problems under non stationary noisy conditions and the final results are strongly dependent on the degradation model, which could be unsuitable in real situations. Finally, NS-MEMLIN and NS-MEMLIN-CPM do avoid the training stereo data requirement of MEMLIN and MEMLIN-CPM without increasing the computational cost in the decoding process.

5. Conclusion

In this work, an unsupervised non-stereo data training process based on ML criterion is presented for Multi-Environment Model-based LInear Normalization (MEMLIN) and MEMLIN with cross-probability model based on GMMs (MEMLIN-CPM). Expectation Maximization (EM) technique is used jointly to Vector Taylor Series (VTS) to provide the ML solution for all the parameters that have to be estimated in the training process. Some results with Spanish SpeechDat Car database show the effective performance of the proposed training procedure (61.9% of average improvement when it is applied over MEMLIN-CPM). Although it is not reached the improvements obtained with the corresponding feature vector normalization techniques with stereo training data, the novel approach provides a more flexible way to obtain the required parameters without a special enrollment by the speakers (no transcription is required and just one far field microphone is needed). As future line, we are studying how to estimate more complex signal models (bias vector transformation plus projection matrix) using the unsupervised non-stereo training framework we have presented in this work.

6. References

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 3, no. 16, pp. 261–291, 1995.

[2] P. Moreno, "Speech recognition in noisy environments," Ph.D. dissertation, ECE Department, Carnegie-Mellon University, Apr. 1996.

[3] J. Droppo and A. Acero, "Maximum mutual information splice transform for seen and unseen conditions," in *Interspeech*, 2005.

[4] 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.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the *EM* algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–21, 1977.

[6] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-performance hmm adaptation with joint compensation of additive and convolutive distortions via vector taylor series," in *ASRU*, 2007.

[7] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "Speechdat-car. a large speech database for automotive environments," in *Proceedings of LREC*, vol. 2. Athens, Greece, June 2000, pp. 895–900.

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