EVALUATION OF THE SPACE DENOISING ALGORITHM ON AURORA2

Christophe Cerisara

INRIA-LORIA
54506 Vandoeuvre-les-Nancy. France
cerisara@loria.fr

Khalid Daoudi

IRIT-CNRS
31062 Toulouse cedex9. France
daoudi@irit.fr

ABSTRACT

Recently we introduced a new and simple denoising algorithm, called SPACE, that yielded promising preliminary results in noise robust speech recognition. SPACE is essentially based on GMM modeling of clean an noisy speech. In this paper, we evaluate the performance of SPACE on Aurora2 and show that they are globally not satisfactory, essentially because the Gaussian correspondence assumption is not verified. We then propose a new training procedure for the GMMs that achieves a better Gaussian correspondence. We further develop a simple adaptation algorithm to handle unknown environments that preserves the Gaussian correspondence. We evaluate the new denoising algorithm on Aurora2. The results show that it outperforms the multistyle models, sometimes significantly, on the three test sets of Aurora2.

1. INTRODUCTION

In real world applications, noise robustness of automatic speech recognition systems is an important issue. Many techniques [1] have been proposed to handle the difficult problem of mismatch between training and application conditions. We introduced in [2] a denoising algorithm, named SPACE, for *Stereo-based Piecewise Affine Compensation for Environments* (in reference to the SPLICE algorithm [3], which can be seen as a special case of SPACE). The first step in SPACE is to model noisy speech by a GMM. Then, a clean speech GMM is learned using a minimum mean square criterion that attempts to make correspondence between pairs clean and noisy Gaussians. In other words, the (ideal) goal is to make a speech GMM-clustering such that the acoustic region modeled by a clean Gaussian is the same as the one modeled by the corresponding noisy Gaussian. A denoiser that depends only on the two GMMs parameters is then conceived. Recognition is then performed using pseudo-clean HMMs trained on the denoised training corpus. Preliminary experiments, reported in [2], showed promising results.

In this paper, we first evaluate the performances of SPACE on Aurora2. We show that they are globally not satisfactory essentially because the Gaussian correspondence assumption is not verified. We then propose a new training procedure based on the joint probability modeling of clean and noisy speech to improve this correspondence. Another weakness of SPACE concerns the MAP adaptation procedure that was proposed in [2] to handle noisy environments that are not seen in the training corpus. Actually MAP adaptation may also affect the correspondence between the GMMs, as it is confirmed by Aurora2 experiments. We propose in this paper a new and simple adaptation algorithm that guarantees to preserve the Gaussian correspondence. We evaluate the new denoising algorithm on Aurora2. The results show that it outperforms the multistyle models, sometimes significantly, on the three test sets of Aurora2.

2. BRIEF DESCRIPTION OF SPACE AND MAP-SPACE

The SPACE algorithm first models the noisy speech y by a mixture of I Gaussians: $P(y) = \sum_{i=1}^{I} \tilde{\alpha}_i \tilde{G}_i(y)$ where $\tilde{G}_i = \mathcal{N}(\tilde{\mu}_i, \tilde{Q}_i)$ is a Gaussian of mean $\tilde{\mu}_i$ and diagonal covariance $\tilde{Q}_i$, and $\tilde{\alpha}_i$ is the prior of $\tilde{G}_i$. Then, it models clean speech x by a mixture of I Gaussians: $P(x) = \sum_{i=1}^{I} \alpha_i G_i(x)$ (where $G_i = \mathcal{N}(\mu_i, Q_i)$ is a Gaussian of mean μ_i and diagonal covariance Q_i) in a way that attempts to make each Gaussian G_i corresponds to the noisy Gaussian $\tilde{G}_i$. Such correspondences yield an indication on how the clean features distribution is related to the noisy one in the acoustic region i. We assume that this relationship is affine in each acoustic region i. One thus obtains the mapping that transforms $y \sim \mathcal{N}(\tilde{\mu}_i, \tilde{Q}_i)$ into $x \sim \mathcal{N}(\mu_i, Q_i)$ as $x = (Q_i \tilde{Q}_i^{-1})^{\frac{1}{2}}(y - \tilde{\mu}_i) + \mu_i$. This mapping is the basis of SPACE, that is, we assume that:

$$E[x|y,i] = (Q_i \tilde{Q}_i^{-1})^{\frac{1}{2}}(y - \tilde{\mu}_i) + \mu_i.$$

Then clean feature estimate is given by:

$$\tilde{x}(y) = \sum_i P(i|y) E[x|y,i] \quad (1)$$

where $P(i|y) = \frac{\tilde{\alpha}_i \tilde{G}_i(y)}{\sum_i \tilde{\alpha}_i \tilde{G}_i(y)}$. In order to estimate the parameters (μ_i, Q_i) of the clean speech GMM from the training stereo data $(x_t, y_t)_{1 \le t \le T}$, we use the Minimum Mean Square Error (MMSE) criterion. The objective function to minimize is: $F \stackrel{\triangle}{=} \sum_t E[(x_t - \tilde{x}(y_t))^2 | y_t]$.

When test observations are given in a new environment, the basic idea of the MAP-SPACE algorithm is to use these observations in a MAP criterion to adapt the initial noisy speech GMM to the new environment. The expectation is that such adaptation would eventually keep correspondence between the initial and the new model parameters. That is, if $P(y) = \sum_{i=1}^{I} \hat{\alpha}_i \hat{G}_i(y)$ is the adapted noisy speech GMM, where $\hat{G}_i = \mathcal{N}(\hat{\mu}_i, \hat{Q}_i)$, then each Gaussian $\hat{G}_i$ corresponds to the Gaussian $\tilde{G}_i$ (and thus to G_i). These adapted Gaussians $\hat{G}_i$ replace the original Gaussians $\tilde{G}_i$ in the denoising algorithm (1).

3. EVALUATION ON AURORA2

3.1. Experimental setup

Aurora2 is a standard corpus to compare noise robust speech recognition algorithms [4]. The training corpus is divided into two parts: clean and noisy training. The noisy training set is composed of clean speech corrupted by 16 different noisy conditions (four noise types at four SNRs from 5 dB to 20 dB). The test corpus is divided into three test sets, corresponding to different mismatch conditions. Each test set is further divided into 28 noisy conditions (four noise types at seven SNRs from -5 dB to clean). The standard HTK scripts are used to parameterize the corpus and to train the baseline and multistyle HMMs.

For each noisy condition of the training corpus, a GMM is trained using the maximum likelihood criterion. The corresponding clean GMM is trained on the corresponding clean sentences using the MMSE criterion. This couple of GMMs is used to denoise this noisy training corpus (using (1)). Finally, the pseudo-clean acoustic HMMs are trained with the HTK scripts on all the denoised sentences of the training corpus.

For each testing condition, we first detect the closest training environment using a two-steps process. First, the SNR is estimated with the following algorithm: for each test sentence, the energy is computed on a sliding window of 64ms length. The window with the highest energy is assumed to represent speech, while the window with the lowest energy is assumed to represent noise. The SNR of each sentence is estimated from the ratio of the energies in both of these windows. The average SNR over all sentences is then computed. The closest corresponding SNR of the training corpus is found: four training conditions match this SNR. In the second step, the four noisy GMMs for these four training conditions are compared, and the one that maximizes the likelihood is chosen. The test corpus is then denoised using the parameters of this GMM and its corresponding clean GMM. After denoising, the test corpus is recognized by the pseudo-clean acoustic models.

3.2. SPACE and MAP-SPACE results

In this section, we evaluate the SPACE algorithm on the test set A of Aurora2. The first row of table 1 shows recognition scores of SPACE as a function of the number I of Gaussians used in the GMMs modeling. The scores are averaged over the four noises of test A and the five SNRs: clean, 20 dB, 15 dB, 10 dB and 5 dB (the average is thus made over 20 environments). The multistyle and the baseline scores are also reported. One observes that the multistyle training always outperforms SPACE, for all choices of I. Moreover, there is no apparent stability in the SPACE behavior when the number of Gaussians varies.

We carried out another experiment where there is an SNR mismatch between the training and testing conditions: the training SNRs range from 5 dB to clean conditions, and the test SNR is at 0 dB. The average scores over the four noise types of test A (at 0 dB) are shown in the second row of table 1. In this scenario, the performance of SPACE is globally not satisfactory (except when $I = 4$) and its behavior is still instable. In such scenario however (SNR mismatch), it is MAP-SPACE that should be used in principle. We do not report here the recognition results of MAP-SPACE [1] because they are very similar to SPACE results, with an increased instability. This is also true on Test B and C.

			SPACE						
	baseline	multistyle	2	4	8	16	32	64	128
match	77.5	95.0	85.3	93.8	94.0	91.4	87.1	87.0	90.6
0 dB	17.4	59.2	38.4	59.6	48.8	31.4	27.3	27.0	27.3

Table 1. Results of SPACE on Aurora2 test A

There may exist different reasons that explain the (relatively) bad results of SPACE and MAP-SPACE. But definitely the most important one is the fact that the Gaussian correspondence hypothesis is not verified. This means, that the MMSE criterion is not the best way to build such correspondences. We thus have to find another way to train clean and noisy GMMs in order to achieve such correspondence. This is the purpose of the next section.

4. JOINT MODELING OF CLEAN AND NOISY SPEECH DISTRIBUTION

4.1. Modeling $P(x, y)$

In order to achieve a better correspondence between clean and noisy Gaussians, we opted for a joint modeling of the clean and noisy speech distribution $P(x, y)$. This is possible because stereo data is available. Formally, we model $P(x, y)$ using a diagonal-covariance GMM:

$$P(x, y) = \sum_{i=1}^{I} \beta_i H_i(x, y) \tag{2}$$

[1] All the detailed results of the experiments reported in this paper can be found at http://www.loria.fr/~cerisara/

where $H_i = \mathcal{N}(m_i, \Sigma_i)$. Obviously we can always write $m_i \triangleq (\mu_i, \tilde{\mu}_i)$ and $\Sigma_i \triangleq \begin{bmatrix} Q_i & 0 \\ 0 & \tilde{Q}_i \end{bmatrix}$ in such way that

$$P(x) = \sum_{i=1}^{I} \beta_i \mathcal{N}(\mu_i, Q_i) \; ; \; P(y) = \sum_{i=1}^{I} \beta_i \mathcal{N}(\tilde{\mu}_i, \tilde{Q}_i).$$

Once these means and covariances are computed, we can then use (1) as our new denoiser. Note that β_i has to be used instead of $\tilde{\alpha}_i$. We call this denoiser SPACE-JM, where JM stands for Joint Modeling.

Of course the choice of a diagonal form for Σ_i is made for implementation simplicity. A zero cross-variance between x and y is definitely not a valid assumption. We use this assumption however only for the clustering purpose to achieve Gaussian correspondence between the clean and noisy GMMs. As we will see in the next section, SPACE-JM leads indeed to much better performance than SPACE.

4.2. Comparison of SPACE and SPACE-JM

In this section, we show a comparison of the SPACE and SPACE-JM algorithms on test set A, in match (figure 1) and mismatch SNR (figure 2) conditions. The setup of SPACE-JM is the same as SPACE, the only change is in the way $(\beta_i, \mu_i, \tilde{\mu}_i, Q_i, \tilde{Q}_i)$ are estimated. In SPACE-JM, (2) is used to estimate these parameters, instead of MMSE for SPACE.

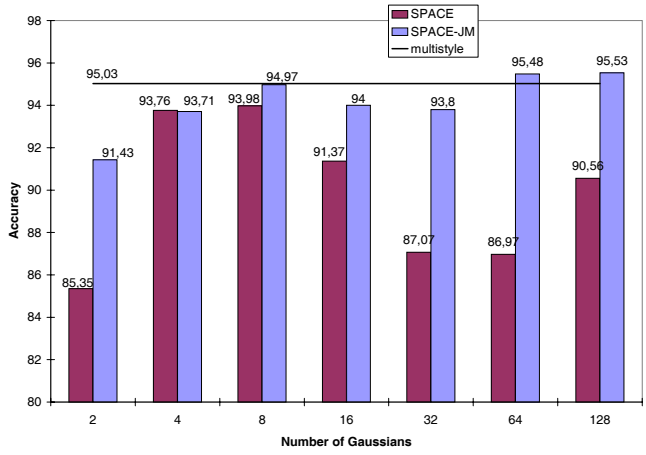

Fig. 1. Average accuracy over the 4 noises and 5 SNRs (5 dB, 10 dB, 15 dB, 20 dB and clean) of aurora2 test set A

In both experiments, SPACE-JM recognition results are much more stable than SPACE results. Furthermore, when the number of Gaussians increase, the accuracy does not globally decrease, as it is the case for SPACE (the best accuracy in both experiments is obtained for 128 Gaussians). More importantly, SPACE-JM slightly outperforms multistyle (for $I = 64, 128$) when there is no SNR mismatch (figure 1) and significantly outperforms multistyle (for $I = 2, 8, 64, 128$) when there is an SNR mismatch (figure 2). This not only suggests that a better Gaussian correspondence is achieved by SPACE-JM, but also that it is robust to SNR change.

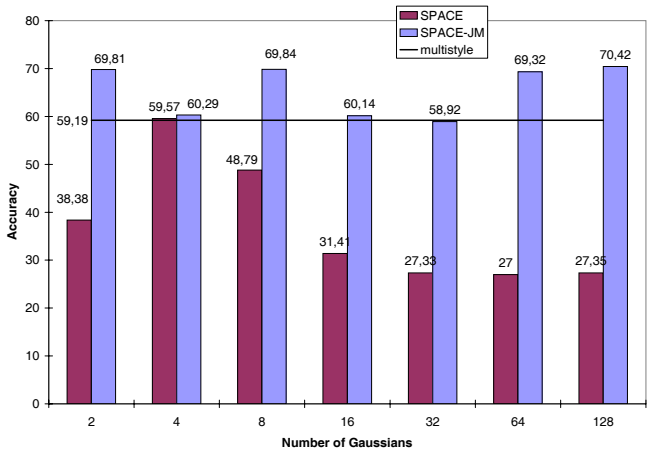

Fig. 2. Average accuracy over the 4 noises of aurora2 test set A at SNR=0 dB

4.3. Adaptation of SPACE-JM to unknown environments

In order to handle unknown noisy conditions that do not belong to the training corpus, the solution that we proposed for SPACE in section 2 is based on MAP adaptation of the noisy GMMs. However, the drawback of this approach is that the correspondence between the clean and noisy GMMs may be affected by the adaptation of the noisy GMMs, and we have seen in the previous experiments that the quality of this correspondence is very important for denoising. Indeed, we carried out experiments on test B and C using SPACE-JM with a MAP adaptation. The recognition scores were even worse than with SPACE-JM alone. This shows that Gaussian correspondence was broken after MAP adaptation. We propose in this section a simple adaptation method for SPACE-JM that guarantees to keep the correspondence between the clean and noisy Gaussians.

For each new test condition, the global mean ν_{test} is first computed by averaging all the observations of the set of sentences for this test condition. Then, the closest training environment is found as described in section 3.1. The global mean ν_{train} of this training environment is also computed.

Adaptation proceeds by adding the global bias $\nu_{test} - \nu_{train}$ to the mean of every Gaussian $\tilde{G}_i = \mathcal{N}(\tilde{\mu}_i, \tilde{Q}_i)$ of this training environment. The adapted noisy Gaussians are thus $\hat{G}_i = \mathcal{N}(\hat{\mu}_i, \tilde{Q}_i)$, where:

$$\hat{\mu}_i = \tilde{\mu}_i + \nu_{test} - \nu_{train} \tag{3}$$

These mean-adapted Gaussians $\hat{G}_i$ replace the original Gaussians $\tilde{G}_i$ in the denoising algorithm (1).

With this adaptation procedure, all the Gaussians of the noisy GMM are globally adapted with the same bias. Hence, their topological relationships are not altered, nor their correspondence with the clean GMM. We refer to this new system as B-SPACE-JM, for *Bias-adapted SPACE-JM*.

4.4. Evaluation of B-SPACE-JM on the three test sets

In this section, we evaluate the B-SPACE-JM algorithm on the three Aurora2 test sets. Figures 3, 4 and 5 respectively present the average recognition accuracy over the whole test sets A, B and C. The best recognition accuracy is obtained by the B-SPACE-JM algorithm on every test set.

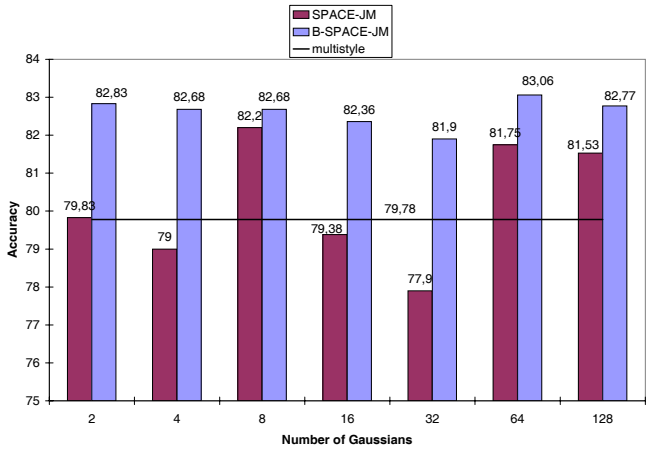

Fig. 3. Average accuracy over all the environments of test A

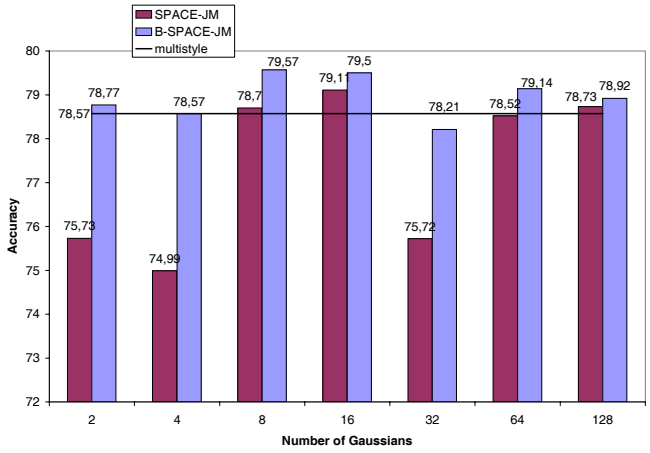

Fig. 4. Average accuracy over all the environments of test B

The good performances of B-SPACE-JM on test set A, where the environments are quite close to the training environments, can be explained by the fact that the adaptation procedure compensates for the errors realized when the closest training environment is estimated. Despite the simplicity of the adaptation procedure, adapting the noisy GMMs brings a clear improvement over the unadapted system in all conditions, particularly in Test C where the improvement is

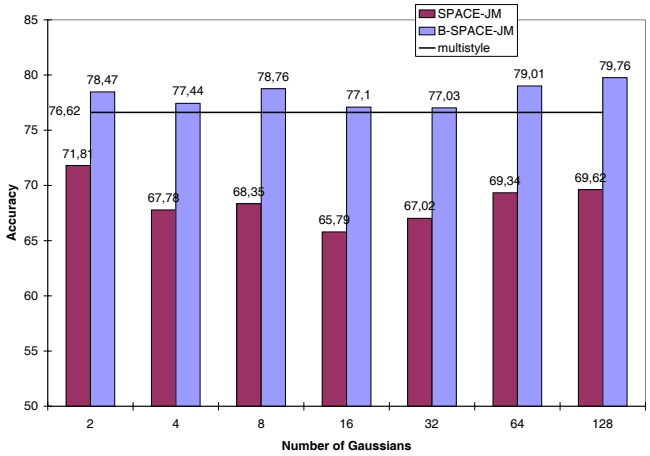

Fig. 5. Average accuracy over all the environments of test C

significant. This confirms the importance of adaptation in our approach. One can also note that the optimal number of Gaussians is dependent on the environment: it is respectively 64, 8 and 128 for tests A, B and C. This may be due to the fact that, although the clean and noisy GMMs are correctly aligned, they do not accurately represent the regions of the acoustic space that are differently affected by noise. On the other hand, good recognition results are consistently obtained for a large number of Gaussians (64 and 128).

5. CONCLUSION

In this work, we proposed a simple but yet effective denoising algorithm for robust speech recognition. The evaluation on Aurora2 shows that this algorithm outperforms the multistyle training and is robust to SNR, noise and channel mismatch. Moreover, many perspectives for further improvements are possible. For instance, adaptation could be improved by transforming the variances of the noisy GMM. Another possibility is to improve the quality of denoising by using sparse covariance matrices for the joint probability distribution. This would allow rotations in the Gaussian mapping in addition to dilatations (only the later is possible when using diagonal covariances). This will be the purpose of future work.

6. REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.

[2] K. Daoudi and C. Cerisara, "The map-space denoising algorithm for noise robust speech recognition," in *Proc. ASRU*, Cancuun, Mexico, 2005.

[3] J. Droppo, L. Deng, and A. Acero, "Evaluation of splice on the aurora 2 and 3 tasks," in *Proc. ICSLP*, 2002.

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