SNR-DEPENDENT FLOORING AND NOISE OVERESTIMATION FOR JOINT APPLICATION OF SPECTRAL SUBTRACTION AND MODEL COMBINATION *

Volker Schless *Fritz Class*

Daimler-Benz AG, Research and Technology, Wilhelm-Runge-Str. 11,
D-89081 Ulm, Germany
e-mail: schless@dbag.ulm.daimlerbenz.com

ABSTRACT

We present an approach to joint application of spectral subtraction (SPS) and model combination (PMC) for speech recognition in noisy environments. Contrary to previous solutions e.g. [2] distortion introduced by SPS is not modeled in PMC. Instead we ensure compatibility of the two methods by adapting parameters of SPS (spectral floor and overestimation factor) according to the present signal-to-noise-ratio (SNR). The scheme leaves the model combination process unchanged which simplifies parameter estimation and reduces computation time. Experiments show significant improvements when using PMC with modified SPS instead of standard SPS.

1. INTRODUCTION

Speech recognition is required in various environments. Besides other fields the use of speech controlled systems in cars has gained much interest in recent years. Here several functions can be accessed by speech very comfortable and with lower distraction of the driver. Examples are speech controlled cellular phones, navigation systems, broadcasting services and others.
Though usage of speech in that kind of environment requires a highly robust recognizing system. Just imagine the noise inside the car when driving on a highway. But it is not enough to adapt the system to high speeds. Also slowly and more rapidly changing environments should be covered when accelerating or breaking is done or a tunnel is entered.
It seems not to be reasonable to collect speech samples for all these situations and train the system with such sort of data. On the one hand this would cause high efforts of time and costs for recording and training on the other hand the system may then be robust in very special noise situations, but may perform poor in low noise situations.
For these reasons it is more efficient to train the system with clean speech samples and adapt it during recognition like in the model combination (PMC) scheme [3]. Alternatively subtraction of stationary noise can be done using the well known method of spectral subtraction (SPS). Recently the implementation of both methods was discussed in [2] with encouraging results.
Unfortunately combination of both methods can't be done straightforward because SPS introduces distortion in the resulting speech signal and PMC relies on undistorted speech. Effects of SPS are even worse when adaptation of the noise model is done during recognition like in [5].
The approach of this paper suggests a new and fast way of using SPS and PMC without expensive modeling of distortion. Instead we try to minimize distortion by adapting parameters of the SPS according to the SNR what makes modification of PMC unnecessary.
The forthcoming sections include a short outline of SPS and PMC. Preliminary experiments are conducted that motivate our approach and lead to its implementation. Finally experiments and results are presented.

2. SPS AND PMC

Model combination and SPS are both methods for robust speech recognition but work differently. Nevertheless they have potential for joint application because they are based on the model of additive noise in the spectral domain:

$$F(i) = S(i) + N(i) \tag{1}$$

$F(i)$ is the input power spectrum of the i-th frequency band, S and N are the clean speech power spectrum and the noise power spectrum, respectively.
Furthermore they act in subsequent processing steps using different techniques. Thus model combination may compensate for the residual noise left in the signal after SPS.
Spectral subtraction works as follows: In speech pauses the noise characteristics are estimated and simultaneously subtracted from the input signal. To enhance the quality of the results and to minimize distortion caused by the mismatch of the actual and the estimated noise power, two parameters are introduced. These are the overestimation factor α and the spectral floor β. So the clean speech spectrum $\hat{S}(i)$ is estimated as follows [1]:

$$\hat{S}(i) = \begin{cases} F(i) - \alpha E\{N(i)\} & F(i) - \alpha E\{N(i)\} > \beta F(i) \\ \beta F(i) & \text{otherwise} \end{cases} \tag{2}$$

Contrary to that the PMC doesn't clean the speech signal but tries to incorporate the noise into the recognition process. Thus it is necessary to estimate a model for the noise

*This work was partly supported by the German Federal Ministry of Education, Science, Research and Technology (BMFT) under Grant No. 01 IV 102 E. The authors are solely responsible for the contents of this publication.

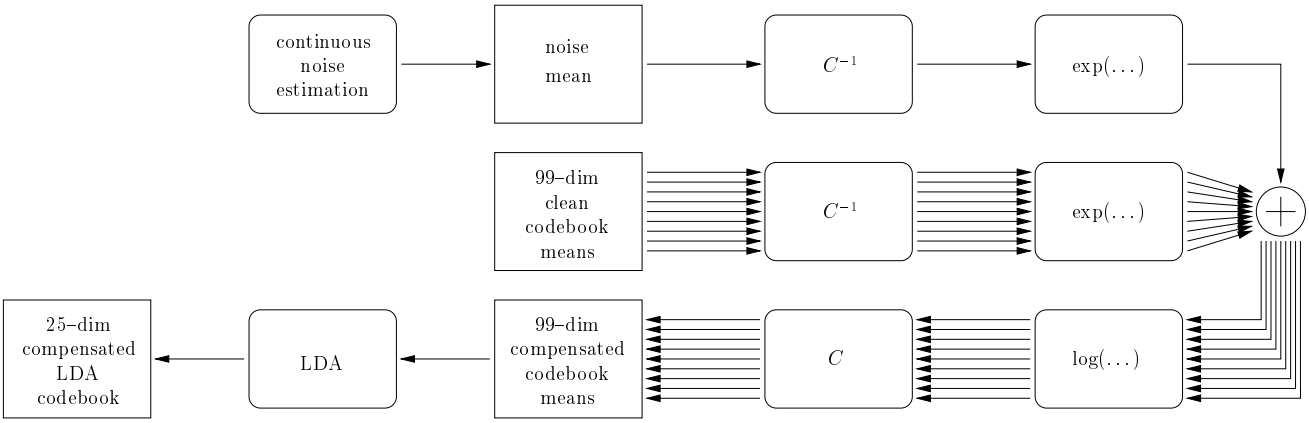

Figure 1: Model combination scheme using LDA–codebook and online noise estimation

and to combine that model with the existing models of clean speech available from training. The theory of model combination is explained in [3]. The extended model combination scheme used in this work is illustrated in Figure 1. Details can be found in [5].

3. MOTIVATION

One of the benefits of this work should be that the model combination process like stated above can be used without modifications. For that purpose we concentrate now on the effects of SPS for subsequent processing.
Normally when SPS is used, training with SPS and a certain amount of noisy speech data is done. However this is not suitable for PMC because model combination requires models of clean speech. Thus a recognizer for clean speech has to be trained. Now the distortion of SPS is not included in the speech models which may cause poor performance when SPS is used during the recognition process. We did some preliminary experiments to examine the effects of SPS on a recognition system trained without SPS. The main goal of these tests is to see what parameter settings are suitable for using SPS in several noise situations. For that purpose we added car noise to clean speech to simulate 6 different noise conditions. The parameters α (overestimation factor) and β (spectral floor) were varied in the interval $\alpha \in [0.0, 0.1, \ldots, 1.4]$ and $\beta \in [0.0, 0.1, \ldots, 1.0]$. In case of $\alpha = 0.0$ or $\beta = 1.0$ no subtraction of noise is done which corresponds to a system without SPS. The results for each driving situation were scaled separately from white (highest recognition accuracy) to black (lowest accuracy) to get an idea of the best parameter set (white area) as plotted in Figure 2.
The figure shows that optimal parameter sets (α,β) are dependent on the current noise situation. It is evident that the noisier the environment the more of the actual noise estimate has to be subtracted. In other words we have to enlarge α and to reduce β.
This may be explained as effects of the implementation of SPS. In reality the noise estimate will only roughly resemble the actual noise mean, because of slight noise changes during speech activity, noise adaptation during speech and the overall non-optimality of speech-pause detection and the noise estimation process. Furthermore the actual noise is always greater or lower than its mean, so too much or too less noise will be subtracted, especially when noise variance is high.
This kind of distortion is harmful if these effects are not taken into account during training. In that case for low noise data SPS should be switched off ($\alpha = 0.0$ or $\beta = 1.0$) while for medium and high noise situations subtraction and overestimation of noise is suitable. Parameter setting should be done to subtract a maximum of noise while minimizing distortion.

4. IMPLEMENTATION

As we have stated in the previous section the subtraction process is governed mainly by the spectral floor and the overestimation factor of SPS. These parameters regulate the amount of noise that is subtracted from the noisy speech signal. Experiments suggest that for each noise level different parameter sets yield optimal performance.
Setting the parameters adaptively according to the noise level leads to undegraded results at high SNR while in low SNR regions the benefits of the noise reduction process are significant. Adapting the SPS to different noise levels was already introduced by Lockwood (nonlinear SPS [4]). This is equivalent to an overestimation factor regulated by a function of the noise estimate.
However like it can be seen in Figure 2 SPS can't be optimized by varying the overestimation factor α alone. Additionally we have to set the spectral floor β in a similar manner and so adapt both overestimation factor and floor for optimal performance.
For that reason we suggest to set both parameters dependent of the current noise situation. Of course the signal-to-noise ratio (SNR) can be used for that purpose. The SNR may be estimated easily by using the noise estimates $E\{N\}$ and the speech estimates $\hat{S}$ of each frequency band:

$$\mathrm{SNR} = 10 * \log_{10} \left(\frac{\sum_i \hat{S}(i)}{\sum_i E\{N(i)\}} \right) \qquad (3)$$

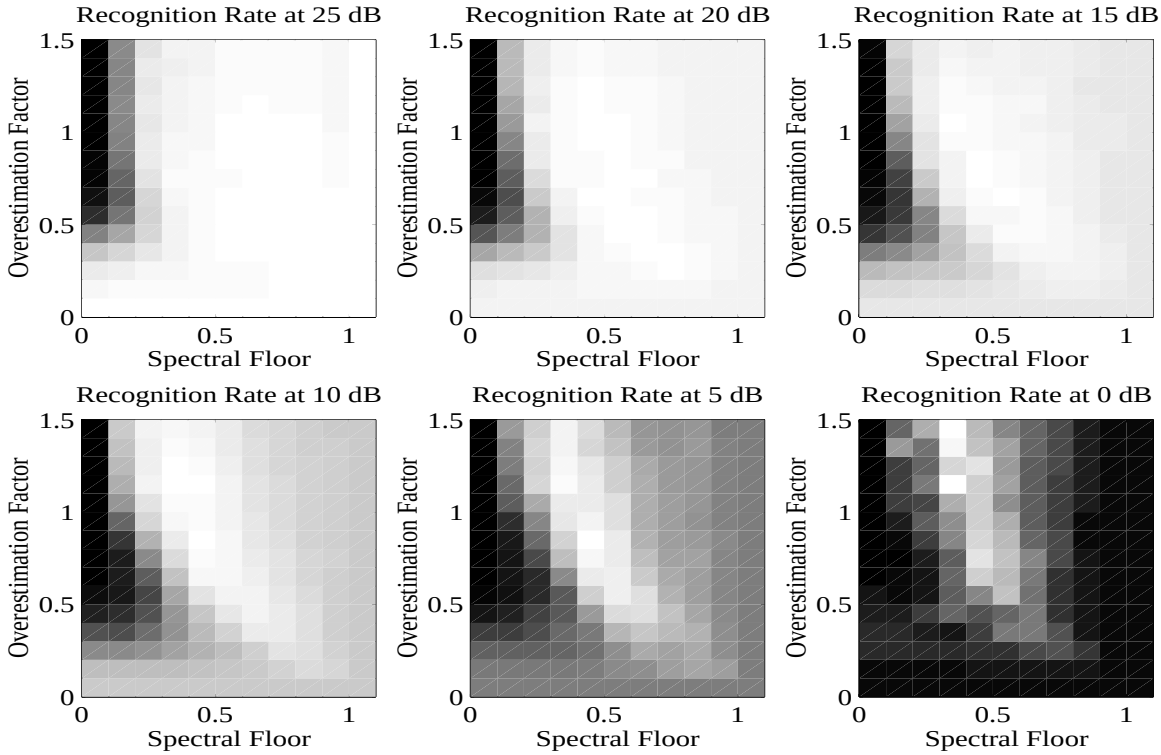

Figure 2: Results of a recognition system trained with clean speech and SPS turned off. Experiments were performed with different parameter sets (α,β) of the SPS for several SNR. Each image was scaled separately from white (highest recognition accuracy) to black (lowest accuracy).

In our implementation the SNR is computed during speech activity only and smoothed over several hundred milliseconds of speech input to get reliable estimates. Now we introduce two functions for regulating the parameters:

$$\alpha = f(\text{SNR}) \quad (4)$$

$$\beta = g(\text{SNR}) \quad (5)$$

To determine suitable functions for setting the parameters further testing was done. In our case application of SPS with PMC has to be optimized. So tests analog to the previous section were performed with SPS and PMC. Due to the additional noise compensation of PMC results are slightly different. In general optimal performance was obtained when subtracting a smaller amount of noise than with SPS alone. Nevertheless results are similar to those illustrated in Figure 2. Examination of the corresponding patterns lead to the following conclusions:
Overestimation was limited to $\alpha \leq 1.0$, because no further improvement could be noticed with increasing α when PMC was active. Also very low flooring didn't improve performance even in high noise situations so β was forced to be greater 0.15.
With this outline and by selecting a linear-type function we suggest to set α and β as formulated in Equations 4 and 5 according to Figure 3. Also non-linear functions have been tested, but there was no evidence that these lead to significantly better results.

5. EXPERIMENTS

Experiments were performed to evaluate the efficiency of the proposed method. For that purpose several test sets were used. The first one includes speech samples from 23 speakers. 100 digit strings containing 3–5 German digits were recorded of each speaker in a standing car (SNR about 28 dB). Recordings of noise in a moving car at 100 km/h and 140 km/h were added to those speech samples. So we obtained test sets of 3 environments (0 km/h, 100 km/h and 140 km/h).
Another test set (called "mixed") not included in training was recorded in a moving car. It contains 1650 digit strings (avg. 3.5 digits per utterance) at different velocities with an average SNR of 8.
Training of the speech recognition system was done with clean speech only in order to use model combination. We used 10 cepstral features plus normalized energy. Vectors of nine subsequent frames were concatenated and transformed with the LDA. PMC was applied to the codebook means before LDA-transformation (see Figure 1).
As can be seen in Figure 4 in case of the car standing (0 km/h) performance of a system trained with clean speech is maintained also with the new SPS. Here the system performs significantly better than without modifications. At 100 km/h both versions of SPS yield comparable results, while at 140 km/h the modified SPS again performs better. Also for the mixed test set the benefits of the new

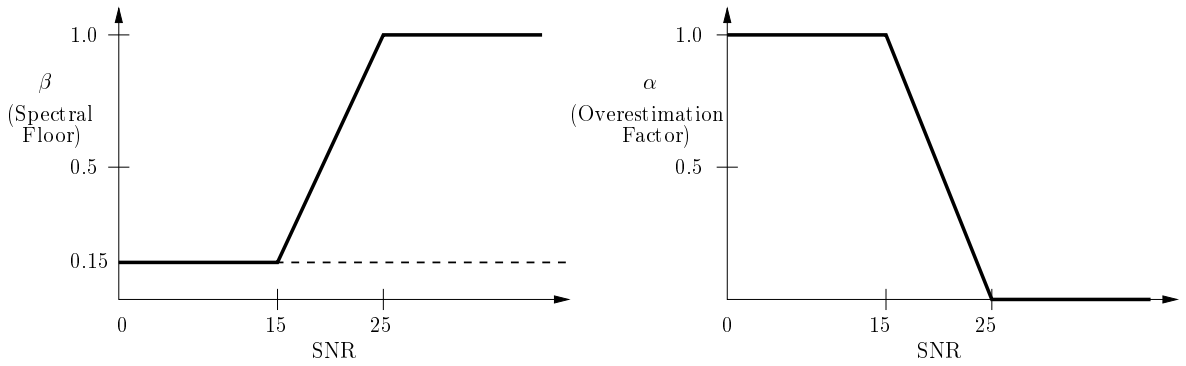

Figure 3: SNR-dependent setting of α and β

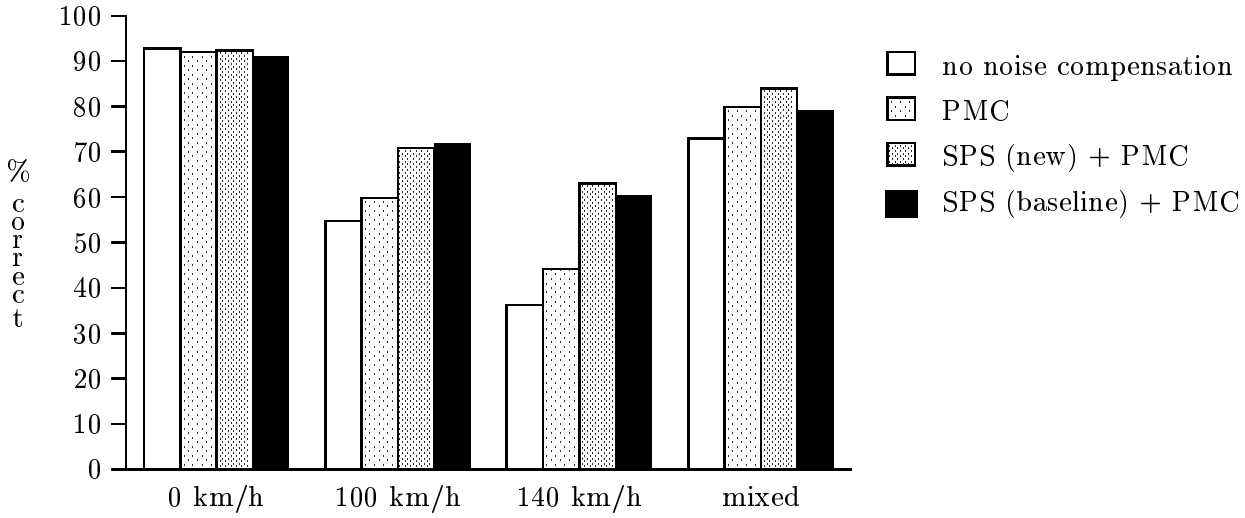

Figure 4: Digit string recognition rates for different recognizers

SPS with PMC are significant.
It may be noticed that noise estimation and adaptation of the noise model for PMC is done after completion of SPS. When using SPS with PMC in that way, training of a noise model before recognition is not appropriate because residual noise varies according to the parameters of the SPS.

6. CONCLUSION

A new scheme of applying both SPS and model combination was introduced, that minimizes distortion of SPS in high SNR regions while improving recognition rate for very noisy speech. This is done by adapting the overestimation factor and the spectral floor according to the SNR. Contrary to previous schemes no consideration of distortion in model combination is necessary. Thus no modification of the PMC scheme is required, which reduces parameters and calculations. Experiments show that the modified SPS yields improved performance compared to a standard SPS-PMC scheme.

REFERENCES

1. M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 208–211, 1979.
2. J. Nolazco Flores and S. Young. Continuous speech recognition in noise using spectral subtraction and HMM adaptation. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 409–412, 1994.
3. M. Gales and S. Young. Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, 4(5):352–359, 1996.
4. P. Lockwood and J. Boudy. Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. *Speech Communication*, 11:215–228, 1992.
5. V. Schless and F. Class. Adaptive model combination for robust speech recognition in car environments. In *Proc. European Conf. on Speech Communication and Technology*, volume 3, pages 1091–1094, 1997.