ADAPTIVE MODEL COMBINATION FOR ROBUST SPEECH RECOGNITION IN CAR ENVIRONMENTS *

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 a new adaptive method for online noise estimation which extends the model combination approach to slowly varying noise conditions. The technique of model combination is reported to improve accuracy in speech recognition without extensive training of noisy speech data. Only training of noise characteristics is needed. However, if the noise characteristics vary over time, calculation of noise parameters once before recognition is not suitable. Therefore the new method of online estimation allows an adaptation to the current noise situation. Furthermore cepstral mean subtraction is added to the model combination scheme. This removes convolutional noise as well. Finally, it is shown how linear discriminant analysis eases handling of dynamical effects for model combination.

1. INTRODUCTION

With the increasing range of applications, automatic speech recognizers are required that dynamically adapt to the environment in which they are used. For example new applications are possible for systems installed in cars. Functions like speech controlled dialing for cellular phones are very useful to reduce the distraction of the driver. Furthermore, speech recognizing systems in cars may be extended to control other parts of the vehicle such as a navigation system, windows, etc.
To perform these tasks the speech recognizer should be able to cope with the effects of a dynamic environment. If the system has been trained with speech samples recorded in quiet surroundings it performs worse in a car than in an environment that more closely resembles training conditions. This is due to varying noises caused e.g. by motor or ventilation. A possible solution to this problem is to perform training in different noise conditions to cope with the mismatch of training and testing environments. However, this significantly increases training time and the recognizer is adjusted only to those noise situations that are included in training. Another method to deal with noisy speech is trying to remove the distortion during preprocessing. This can be done with the well known technique of spectral subtraction for example [1]. Another possibility is to train speech and noise separately and add them during the recognition process using model combination [3]. This technique has been shown to increase recognition capabilities in noisy environments with training limited to clean speech data and a suitable noise model [11, 8]. In case of varying noises training of different kinds of noises and choosing a suitable noise model during the recognition process is necessary. Here the problem arises that it is not possible to get noise samples for every possible situation and even if we could do this, there exists the need to classify the noise to one of the trained noise classes. Additionally, to calculate the appropriate combination weights for speech and noise, an estimation of the SNR is needed. To overcome these problems, an approach to perform an online estimation of the noise characteristics is presented. Secondly we show how to use linear discriminant analysis (LDA) with the model combination technique to obtain a dimensionality–reduced codebook that includes all of the dynamic effects (like Δ and $\Delta\Delta$ features). Finally, the implementation of cepstral mean subtraction to be used with the adaptive model combination is described.

2. ADAPTIVE MODEL COMBINATION

The theory of model combination is explained in [3]. The basic idea is to transform means and covariances of each codebook class and the characteristics of noise from the cepstral to the linear spectral domain. After that mean vectors and covariance matrices of speech and noise are added. Then a transformation of the combined means and covariances back to the cepstral domain is carried out.
The following steps are necessary: First the transformation of mean vectors $\boldsymbol{\mu}$ and covariance matrices $\boldsymbol{\Sigma}$ from the cepstral domain to the logarithm domain is done using the inverse cosine transformation matrix $\boldsymbol{C}^{-1}$.

$$\boldsymbol{\mu}^{log} = \boldsymbol{C}^{-1}\boldsymbol{\mu}^{cep} \quad (1)$$

$$\boldsymbol{\Sigma}^{log} = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}^{cep}\boldsymbol{C} \quad (2)$$

The superscripts denote the domain of means and variances. Now piecewise calculation of the linear elements is carried out.

$$\mu_i^{lin} = \exp\left(\mu_i^{log} + \frac{\sigma_{ii}^{log}}{2}\right) \quad (3)$$

*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.

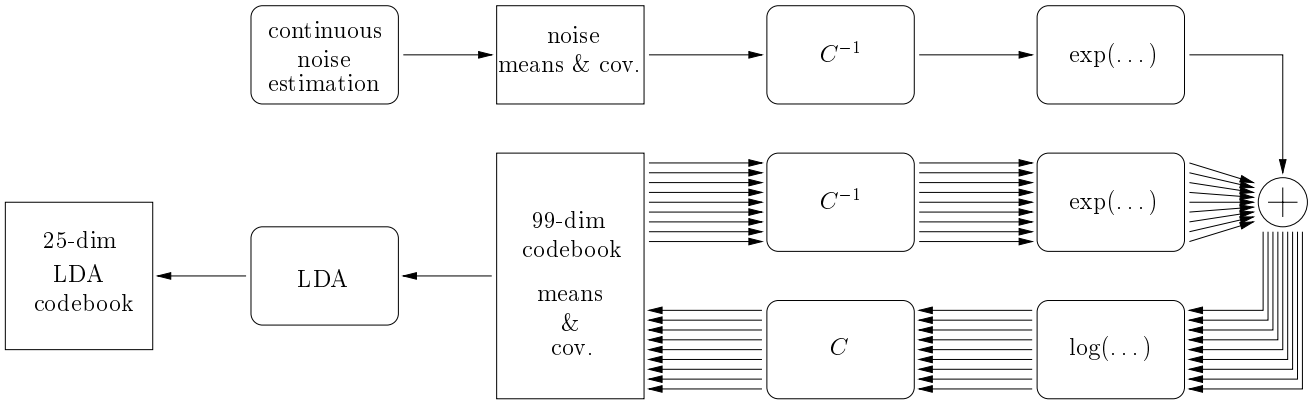

Figure 1: Model combination scheme using LDA–codebook and online noise estimation

$$\sigma_{ij}^{lin} = \mu_i^{lin}\mu_j^{lin}(\exp(\sigma_{ij}^{log}) - 1) \quad (4)$$

The characteristics of noise and speech can now be added using a weighting factor g that depends on the signal to noise ratio (SNR).

$$\boldsymbol{\mu}^{lin} = g\boldsymbol{\mu}_s^{lin} + \boldsymbol{\mu}_n^{lin} \quad (5)$$

$$\boldsymbol{\Sigma}^{lin} = g^2\boldsymbol{\Sigma}_s^{lin} + \boldsymbol{\Sigma}_n^{lin} \quad (6)$$

After combination the retransformation to the logarithm domain has to be performed.

$$\mu_i^{log} = \log(\mu_i^{lin}) - \frac{1}{2}\log\left(\frac{\sigma_{ii}^{lin}}{\mu_i^{lin}\mu_i^{lin}} + 1\right) \quad (7)$$

$$\Sigma_{ij}^{log} = \log\left(\frac{\sigma_{ij}^{lin}}{\mu_i^{lin}\mu_j^{lin}} + 1\right) \quad (8)$$

Finally, we get back to the cepstral domain by applying the cosine transformation.

$$\boldsymbol{\mu}^{cep} = \boldsymbol{C}\boldsymbol{\mu}^{log} \quad (9)$$

$$\boldsymbol{\Sigma}^{cep} = \boldsymbol{C}\boldsymbol{\Sigma}^{log}\boldsymbol{C}^{-1} \quad (10)$$

After these calculations we obtain a codebook that is adapted to the noise parameters under the assumption of the noise being additive in the linear domain.

Instead of training of certain noise characteristics before the recognition process, as it is done in previous work [3, 11], we apply an adaptive method to maintain a flexible system. To estimate noise parameters we use non–speech frames of the input data in the cepstral domain. For this purpose only frames in a small interval around the energy minima of each phrase are considered. Each time before model combination is carried out, mean and covariance of these noise frames are calculated. This can be done between consecutive utterances. After collecting a certain number of frames, the adjustment to variations of noise characteristics is done. Thus, depending on the actual situation the noise characteristics to be combined are estimated from non–speech segments, successively updated and added to the speech codebook.

3. LDA–TRANSFORMATION

Recently, experiments with combination of noise and speech for static, dynamic and velocity features have been done. Using time derivatives resulted in a further reduction of error rate [4]. Contrary to this article we propose a method that is a straightforward extension to the combination of static features.

The information of Δ and $\Delta\Delta$ features is included in the preceding and succeeding feature vectors. Thus, in order to model the time characteristics of speech, we concatenate static features of the four preceding and four succeeding vectors to one feature vector. 10 cepstral coefficients plus one energy coefficient are used as static features. With a time window of 9 frames we obtain a feature vector of dimension 99.

The LDA works well for dimensionality reduction and better class separation [2, 6, 9]. After transforming the 99–dimensional feature vector with the LDA–matrix we just take the first 25 components and build up a smaller codebook from these vectors.

For model combination we have to go back to the linear domain. Because of the reduction from 99 to 25 dimensions we have lost too much information to maintain a proper retransformation from the LDA domain to the cepstral domain. Therefore it is necessary to save a copy of the 99–dimensional codebook, which can be used for transformation to the linear domain and for combination with the noise coefficients.

The transformation of the 99–dimensional codebook is done as follows: Because each mean vector of the codebook has been generated by 9 static feature vectors, we can split it into its components. The 9 parts are treated separately as it is done in original model combination. The energy information is left unchanged. Because of a normalization procedure during preprocessing the energy value of speech frames remains unchanged independent of the present noise level. Figure 1 illustrates the whole process. Every component is transformed to the linear domain and combined with the estimated noise characteristics. Back in the cepstral domain the components are concatenated again. Now the LDA–transformation is used to maintain an updated LDA–codebook. Thus, we have derived a noise adapted 25–dimensional LDA–codebook from a 99–dimensional codebook that has been trained with clean data only.

4. CEPSTRAL MEAN SUBTRACTION

Another positive side effect of the online estimation of noise in the cepstral domain is the possibility to use cepstral mean subtraction. CMS is a method to reduce speaker dependency and distortion [2, 5]. It can be shown that the long–term average of the cepstral vectors represents the channel characteristics [7]. To remove these effects we continuously calculate the mean of the cepstral vectors and subtract it from the actual feature vector. The adaptation speed of the cepstral mean is determined by time constants of an exponential window.

Model combination uses a system trained with clean speech. Therefore the cepstral mean is estimated from clean speech samples only. If the noise characteristics are trained offline, the cepstral mean of the speech during training does not contain the mean of the noise cepstral coefficients.

During recognition the cepstral mean will adjust to the cepstral features of noisy speech. Consequently, the residual noise characteristics diverge from the noise characteristics that were trained separately before and the combined codebook does not match the actual noise condition.

However, with our method of adaptive noise estimation in the cepstral domain the noise level is set correctly to the current situation. This is because the cepstral subtraction process is done before noise estimation and so the codebook is adapted only to the remaining noise level.

5. EXPERIMENTS AND RESULTS

Experiments were performed to evaluate the efficiency of the proposed methods. For that purpose we took speech samples from 60 speakers (33 male and 27 female). 100 digit strings containing 3–5 German digits were recorded for each speaker in a standing car (SNR about 28 dB). For testing, recordings of noise in a moving car at 100 km/h and 140 km/h were added to the speech samples. This results in a realistic SNR of 12 dB and 8 dB, respectively. The evaluation was done with utterances of digit strings from 6 speakers (3 male and 3 female) yielding 600 samples for each of the 3 environments (0 km/h, 100 km/h and 140 km/h). Testing utterances are not included in the training set.

The environments are tested one after another without restart. Estimation of noise characteristics is done continuously for the different environments. So adaptation to each new environment is necessary.

The system is based on semi–continuous HMM's of subword units. For the Gaussian distribution full covariance matrices are used. The HMM's consist of 3 emitting states with a loop for each state and a one–state skip. Feature vectors are generated every 10 ms. For further details see [2].

First a system has been trained without CMS and LDA. For recognition we used 3 codebooks based on static, dynamic and energy features respectively. Model combination is performed on the static codebook only.

Without noise compensation the system performed poor in noisy conditions (see line 2 in Table 1). This was compared to model combination with continuous noise estimation added to the recognizer. At the beginning of each new phrase codebook combination with the previously estimated noise characteristics was carried out. Because transformation and combination of covariances is very time consuming we also performed tests with combination of means only. Then the method becomes quite similar to the state–based Wiener filtering as described in [10].

The combination of the means raises the total string recognition rate from 26.9% to 40.3%. Additional 2% is achieved by combining both means and variances. It can be seen for the different levels of noise that especially the performance at 140 km/h increased drastically from 2.5% to 26.3% and 29.3% as shown in lines 3 and 4 of Table 1.

combination method	recognition rate at velocity (km/h)			
	total	0	100	140
no combination	26.9%	59.3%	18.6%	2.5%
means	40.3%	60.0%	34.6%	26.3%
means + variances	42.4%	59.7%	37.9%	29.3%

Table 1: Recognition results for digit strings with different combination methods

Now a system including cepstral mean subtraction will be examined. The other parameters of the recognizer remain unchanged. As can be seen in Table 2 the recognition rate for the system raises from 26.9% to 42.7% using the CMS during training and testing. The results for the mean combination and the mean + variance combination improve as well. However the model combination yields smaller improvement now than without CMS. Omitting variance combination caused a loss of performance of less than 1% while saving a great amount of computation time.

combination method	recognition rate at velocity (km/h)			
	total	0	100	140
no combination	42.7%	76.2%	34.2%	17.4%
means	47.7%	74.3%	42.2%	26.5%
means + variances	48.4%	74.3%	43.5%	27.1%

Table 2: Recognition results for digit strings with different combination methods including CMS

Finally LDA–transformation of the codebook is used for compensating the complete codebook (not only the static codebook as it was done so far). The CMS is still included because it showed significant improvement before. Contrary to the above experiments compensation of variances is not carried out since it seems impractical in real–time applications. This is due to the additional computation time that would be necessary for transforming the codebook matrices to the LDA–domain and the increase of dimensionality compared to the static codebook used before. Moreover it showed no significant improvement to the sole compensation of means when including the CMS.

Results of the new recognizer can be seen in Table 3. The rates of both systems increase applying the LDA–codebook. For no combination the system performs about 3%, for mean combination about 5% better. Adding mean combination to the LDA system yields significant improvement of close to 8%.

method	recognition rate at velocity (km/h)			
	total	0	100	140
no combination	45.5%	78.2%	36.8%	21.4%
mean combination	53.4%	75.2%	48.1%	36.7%
spectral subtraction	47.9%	62.7%	41.7%	39.2%

Table 3: Compensated means of LDA–codebook compared to a system without noise compensation and a system with spectral subtraction

To relate the effectiveness of adaptive model combination to a different noise reduction scheme we performed another training. In a preprocessing step spectral subtraction (SPS) was applied during training and testing [1]. To obtain suitable noisy speech data for SPS training, both car noises were added to the clean speech samples, tripling the amount of training samples.
Although training of noisy data was included the results are worse than the results of adaptive model combination. Due to the distortion that comes with the SPS, recognition rate for clean speech data is poor. At 100 km/h results are better than without compensation but worse than model combination. Only at 140 km/h results for SPS slightly outperform model combination. Similar results for both methods for very noisy speech may be partly due to a significant amount of wrong segmentation. This problem in continuous speech recognition has to be further investigated.
Finally the computation times for the proposed methods will be compared. As can be seen in Table 4 the requirements differ significantly. Combining means in the static codebook took 0.05 seconds on a DEC Alpha 300 MHz. Including full variances of the static codebook requires considerably more time (0.17 sec). Applying adaptive model combination to the whole LDA codebook (information of temporal derivatives included) is more efficient.

mean combination	0.05 sec
mean + variance combination	0.17 sec
mean combination + LDA transformation	0.09 sec

Table 4: Runtime for different combination schemes of a codebook (static or LDA) on a DEC Alpha 300 MHz

6. CONCLUSION

In this contribution we proposed a new adaptive approach to model combination. This is essential in slowly varying noise conditions such as car environments, because training of noise characteristics can not be performed for all possible situations. By estimating noise parameters in the cepstral domain, cepstral mean subtraction can be added to reduce convolutional noise. Additionally, application of LDA–transformation was introduced to reduce dimensionality in the context of model combination. All of the methods yield significant improvements for the model combination scheme. Further tests show that adaptive model combination compares well with spectral subtraction.
Future work will include experiments with speech data recorded in moving cars. Also methods for segmentation of noisy speech are under investigation.

REFERENCES

1. S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2):113–120, 1979.
2. F. Class, A. Kaltenmeier, and P. Regel-Brietzmann. Optimization of an HMM-based continuous speech recognizer. In *Proc. European Conf. on Speech Communication and Technology*, pages 803–806, Berlin, Germany, 1993.
3. M.J.F. Gales and S.J. Young. Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, 12:231–239, 1993.
4. M.J.F. Gales and S.J. Young. Parallel model combination on a noise corrupted resource management task. In *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 255–258, 1994.
5. S. Gupta, F. Soong, and R. Haimi-Cohen. High–accuracy connected digit recognition for mobile applications. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 57–60, 1996.
6. A. Hauenstein and E. Marschall. Methods for improved speech recognition over telephone lines. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 425–428, 1995.
7. C. Mokbel, D. Jouvet, and J. Monné. Deconvolution of telephone line effects for speech recognition. *Speech Communication*, 19:185–196, 1996.
8. S. Nakamura, T. Takiguchi, and K. Shikano. Noise and room acoustics distorted speech recognition by HMM composition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 69–72, 1996.
9. O. Siohan. On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 125–128, 1995.
10. S.V. Vaseghi and B.P. Milner. Noise–adaptive hidden Markov models based on Wiener filters. In *Proc. European Conf. on Speech Communication and Technology*, volume 2, pages 1023–1026, 1993.
11. R. Yang and P. Haavisto. An improved noise compensation algorithm for speech recognition in noise. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 49–52, 1996.