# CEPSTRAL NOISE SUBTRACTION FOR ROBUST AUTOMATIC SPEECH RECOGNITION

*Robert Rehr and Timo Gerkmann*

Speech Signal Processing Group, Department of Medical Physics and Acoustics
Cluster of Excellence Hearing4all, University of Oldenburg, Germany
robert.rehr@uni-oldenburg.de, timo.gerkmann@uni-oldenburg.de

## ABSTRACT

The robustness of speech recognizers towards noise can be increased by normalizing the statistical moments of the Mel-frequency cepstral coefficients (MFCCs), e. g. by using cepstral mean normalization (CMN) or cepstral mean and variance normalization (CMVN). The necessary statistics are estimated over a long time window and often, a complete utterance is chosen. Consequently, changes in the background noise can only be tracked to a limited extent which poses a restriction to the performance gain that can be achieved by these techniques. In contrast, algorithms recently developed for single-channel speech enhancement allow to track the background noise quickly. In this paper, we aim at combining speech enhancement techniques and feature normalization methods. For this, we propose to transform an estimate of the noise power spectral density to the MFCC domain, where we subtract it from the noisy MFCCs. This is followed by a conventional CMVN. For background noises that are too instationary for CMVN but can be tracked by the noise estimator, we show that this processing leads to an improvement in comparison to the sole application of CMVN. The observed performance gain emerges especially in low signal-to-noise-ratios.

***Index Terms***— automatic speech recognition, cepstral analysis, feature normalization, noise robustness, speech enhancement

## 1. INTRODUCTION

The term automatic speech recognition (ASR) describes the process of transcribing speech utterances represented as acoustic wave forms to written words. Nowadays, ASR is used in many applications, e. g. for interacting with mobile devices or home-entertainment systems [1]. Over the last years, many methods and techniques have been described in the literature to make ASR more robust against acoustic influences such as noise and reverberation [2]. One approach for improving the performance of ASR systems is to enhance the feature values extracted from the noisy input signal before the data is statistically modeled, e. g. using hidden Markov models (HMMs) and Gaussian mixture models (GMMs). Such an enhancement can be achieved by normalizing the statistical moments of the feature values. Thus, cepstral mean normalization (CMN) [3] or cepstral mean and variance normalization (CMVN) [4] have been used in many applications as these techniques have proven to give better recognition results in various environments. As shown in [3, 5], the beneficial effect of CMN can be explained by its capabilities to reduce differences between the test and the training data caused by channel distortions and colorizations. Furthermore, in [5], it is discussed that CMN is also able to reduce differences in the feature representation between speakers and can also partly reduce the detrimental influence of the background noise. For CMVN, however, there is no association with a reduction of a specific distortion [5]. However, it can be associated with restoring the temporal diversity of the features which may be reduced due to additive noise which can also be interpreted as a reduction of the feature variance [6]. The general idea of feature normalization has also been taken further, e. g. in [7], where the distribution of the noisy input features is fitted to a target distribution. This procedure normalizes all moments of the input data simultaneously and is also known as histogram equalization.

Even though CMVN offers a simple way to increase the noise robustness of speech recognizers, it is not able to track fast changes in the time-variant noise components as the statistics for the normalization are typically estimated over a longer time-period. In many cases, a whole utterance is chosen. Due to these shortcomings, the performance gain achieved by CMVN is limited. Recent research on single-channel speech enhancement, however, has brought up technologies that allow to track the background noise quickly, e. g. [8]. Therefore, combining these techniques with CMVN may increase the robustness of the recognition towards noise. Unfortunately, enhancing the input signal using state-of-the-art noise reduction algorithms in combination with CMVN, often does not result in an additional performance gain compared to the sole application of CMVN.

Consequently, we investigated other options for including speech enhancement technologies for improving the noise robustness of ASR in combination with CMVN. We propose a novel method which subtracts the Mel-frequency cepstral coefficients (MFCCs) [9] of the background noise from the MFCC vector of the noisy input signal. Subsequently, the processed features are normalized using CMVN. Subtracting the noise component is equivalent to a whitening of the background noise in the Mel spectral domain. We will show that the proposed enhancement technique leads to higher recognition rates compared to the sole application of CMVN if the background noise is too instationary to be compensated by the normalization but can be tracked by a noise estimator. In these cases, the error rate can be reduced – especially in low signal-to-noise-ratio (SNR) conditions. In our experiments we use a state-of-the-art noise estimator for which we employ the algorithm described in [8].

The paper is structured as follows. First, we give a detailed description of the proposed method in Section 2, which is followed by a brief analysis in Section 3. The setup for the experimental evaluation of the proposed method is presented in Section 4 and the results are shown in Section 5.

## 2. PROPOSED METHOD

In this section, we propose to normalize cepstral features based on an estimate of the noise power spectral density (PSD) obtained using a

This work was funded by the PhD program “Signals and Cognition”, and the DFG Research Grant GE2538/2-1.

state-of-the-art noise tracker. We assume that the speech signal $s[n]$ is corrupted by an interfering noise $d[n]$ and that only the noisy input signal $x[n]$ is available. The time domain signal $x[n]$ is split into overlapping blocks which are transformed to the frequency domain using a discrete Fourier transform (DFT) after applying a Hamming window. From this, we compute the periodogram of the input signal $|X[k,\ell]|^2$ where $k$ denotes the frequency index and $\ell$ the block index. This quantity is used to obtain an estimate of the background noise PSD $\hat{\sigma}_d^2[k,\ell]$ by employing the speech presence probability (SPP) based estimator described in [8]. With the coefficients of the $m$th triangular filter $R_m[k]$, the spectral quantities are transformed to MFCCs [9] using

$$X'[m,\ell] = \sum_{k=0}^{N-1} R_m[k]\,|X[k,\ell]|^2\,, \tag{1}$$

$$\mathcal{X}[q,\ell] = \sum_{m=0}^{M-1} \cos\left(\frac{\pi(2m+1)q}{2M}\right) \log\left(X'[m,\ell]\right)\,, \tag{2}$$

where $q = 0,\dots,Q-1$. Here, $N$ and $M$ denote the number of DFT coefficients and Mel filters, respectively, while $Q$ is the number of MFCCs.

Finally, we obtain a processed version of the feature values by subtracting the noise estimate in the cepstral domain via

$$\tilde{\mathcal{X}}[q,\ell] = \mathcal{X}[q,\ell] - \hat{\mathcal{D}}[q,\ell], \tag{3}$$

where $\hat{\mathcal{D}}[q,\ell]$ is the background noise estimate $\hat{\sigma}_d^2[k,\ell]$ transformed to the MFCC domain using (1) and (2). As the subtraction in the logarithmic domain equals a division the linear domain, this operation leads to a whitening of the background noise in the Mel filterbank representation. Before the enhanced features are fed to the speech recognizer they are normalized by a CMVN which is given by [4]

$$\mathcal{M}[q] = \frac{1}{L}\sum_{\ell=0}^{L-1} \tilde{\mathcal{X}}[q,\ell], \tag{4}$$

$$\mathcal{V}[q] = \frac{1}{L-1}\sum_{\ell=0}^{L-1} \left(\tilde{\mathcal{X}}[q,\ell] - \mathcal{M}[q]\right)^2\,, \tag{5}$$

$$\bar{\mathcal{X}}[q,\ell] = \frac{\tilde{\mathcal{X}}[q,\ell] - \mathcal{M}[q]}{\sqrt{\mathcal{V}[q]}}, \tag{6}$$

where $L$ is the number of blocks in a single utterance.

## 3. ANALYSIS AND COMPARISON

Before the experimental results are presented, a short analysis of the proposed preprocessing and its behavior is given. Here, we will point out the differences and possible advantages and disadvantages in comparison to the sole application of CMVN.

For explaining the differences between the two preprocessing strategies, namely CMVN and the method proposed in Section 2, we will make use of an example where a speech utterance has been corrupted by a modulated white noise. The modulation of the background noise is created by applying a time-variant filter function which periodically changes its frequency response from a high-pass to a low-pass characteristic and vice versa. Thus, the modulation of the background noise is frequency dependent which can be observed in the spectrogram of the corrupted speech utterance which is shown in the lower panel of Figure 1. The SNR is set to $-5$ dB.

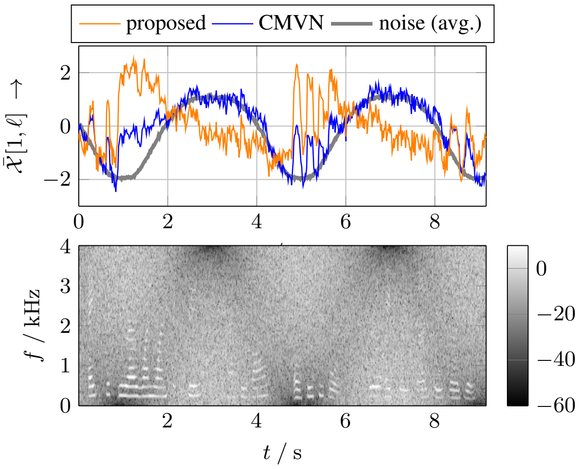

**Fig. 1**: Upper panel: Time course of the second MFCC of an utterance corrupted by frequency-dependent modulated white noise at an SNRs of -5 dB after applying CMVN and the proposed preprocessing strategy. Additionally shown: estimate of the noise power spectral density transformed to the MFCC domain. Lower panel: Spectrogram of the noisy speech signal ($t$: time, $\mathcal{X}[q,\ell]$: $q$th MFCC at frame $\ell$).

As the second MFCC characterizes the spectral tilt, meaning the balance between high and low frequencies, the value of this coefficient will vary depending on the modulation imprinted on the background noise. The upper panel of Figure 1 shows the corresponding time course which results after applying CMVN and the proposed method to the noisy speech feature. Additionally, the ensemble average of the background noise is displayed. It is obtained by averaging the squared magnitude spectra of multiple realizations of the same modulated noise followed by the transformation to the MFCC domain. In other words, this curve represents an estimate of the noise power spectral density transformed to the MFCC domain. For a better comparison, this average is normalized using the means and variances which have been computed for the feature values of the noisy speech MFCCs.

Figure 1 shows that the feature values obtained using CMVN still follow the modulation of the noise. From this, it can be concluded that CMVN has no effect on the time-dependent variations. For the proposed method, however, it is possible to see that the modulation of the feature values is reduced which results in a time-course which is considerably different compared to the feature values obtained by applying CMVN.

The proposed method reduces the impact of the noise modulations in the MFCC domain, and effectively results in a whitening of the noise. However, this whitening will also affect the input speech, which may have a negative impact on the recognition accuracy of the ASR system. However, our experiments, where we used the SPP-based noise estimator to obtain $\hat{\mathcal{D}}[q,\ell]$, show that this processing does not lead to a severe degradation of the recognition accuracy. If trained on multi condition data, it may be possible for the recognition system to learn the disturbances which reduces the influence of them on the recognition rate. However, no large deteriorations are observed if the recognizer is trained on clean speech data. This indicates that the ASR performance is presumably not impaired by the proposed preprocessing.

## 4. EXPERIMENTAL SETUP

In this section, we describe the experimental setup for evaluating the proposed feature enhancement method and how it is compared to other feature normalization techniques such as CMVN. Additionally, we will compare these two methods to a state-of-the-art noise reduction scheme which is combined with CMVN. An important difference to our proposed approach is that here the noise compensation is obtained in the short-time DFT domain. In contrast, in our proposed method the noise compensation is achieved directly in the MFCC domain. CMVN as described in (4) – (6) without further preprocessing serves as baseline in our comparison.

The sole application of CMVN is compared to CMVN combined with the proposed feature enhancement, as described in Section 2. For estimating the background noise, we employ the SPP based noise estimator which has been presented in [8]. This algorithm updates its noise estimate using a SPP soft-decision mask. In our experiments, we decided to deactivate the lock-up protection which tries to avoid stagnation of the algorithm if the SPP is close to 100 % for a longer time period. In situations where no or only little noise is present, this protection can cause the speech power to leak into the noise estimate which affects the recognition results.

Lastly, a state-of-the-art single-channel speech enhancement scheme is included as preprocessing. Here, the features are extracted from an estimate of the clean speech signal $\hat{s}[n]$. After that, the feature values are normalized by a CMVN. The employed noise reduction algorithm uses the Wiener filter to suppress the background noise. It is given by

$$G[k,\ell] = \max\left(\frac{\sigma_s^2[k,\ell]}{\sigma_s^2[k,\ell]+\sigma_d^2[k,\ell]}, G_{\text{min}}\right). \tag{7}$$

An estimate of the background noise $\hat{\sigma}_d^2[k,\ell]$ is obtained by the SPP based estimator proposed in [8]. The speech PSD $\hat{\sigma}_s^2[k,\ell]$ is estimated using temporal cepstrum smoothing which has been described in [10]. This method has been chosen as it has proven to perform better in ASR applications than other state-of-the-art techniques, e. g. the decision-directed approach [11]. Further, a spectral floor $G_{\text{min}}$ is introduced in order to reduce the musical-noise further. This type of artifact has been found to be detrimental for the ASR performance, e. g. in [11]. The parameter $G_{\text{min}}$ is set to $-10$ dB.

Within our experiments, MFCCs, as described in (2), serve as features. We extract the MFCCs using a 23 band Mel filter bank which is limited to the frequency range between 64 Hz and 4 kHz. After applying the discrete cosine transform, we use the lowest 13 coefficients as features where we include the 0th cepstral coefficient. Further, the commonly used first and second order delta derivatives are extracted and appended to the feature vector which results in a 39-dimensional representation of the input signal. For all processing strategies a block length of 32 ms and an overlap of 50 % is chosen. Before a block is transformed to the Fourier domain, it is weighted by a Hamming window. The computation of the empirical means and variances required for the CMVN is utterance based.

For the evaluation, an HTK [12] based speech recognizer and the Aurora2 database [13] is used. For recognizing the speech utterances, a conventional ASR system based on GMMs and HMMs is utilized. The parameters which have been used for the reference recognizer of the Aurora2 experiment [13] are also employed in this evaluation. Consequently, the HMMs consist of 16 states per word and only left-to-right transitions without skips are allowed. The GMMs consist of three mixtures and the covariance matrices are restricted to be diagonal. The ASR system is trained on the clean and the multi-condition training set which are given in the Aurora2 corpus [13]. The results are presented as average over the three test sets of the audio database. Further, the feature processing strategies are also evaluated on noise types which are not part of the Aurora2 database. These noise types change their spectral shape over time which can be tracked by the SPP based estimator but cannot be followed by a conventional CMVN. These additional background noises include seashore noise[1] and traffic noise[2]. For the traffic noise, the first two seconds have been removed in our experiments. In all evaluations, the additional noise types are excluded from the multi condition training set.

As comparison criterion, we employ the recognition accuracy which is given by [12]

$$A = \frac{W - E_D - E_S - E_I}{W}, \tag{8}$$

where $W$ is the total number of words and $E_D$, $E_S$ and $E_I$ denote the number of deletion errors, substitution errors and insertion errors, respectively.

## 5. RESULTS

In this section, the results of the experiments described in the previous section are presented.

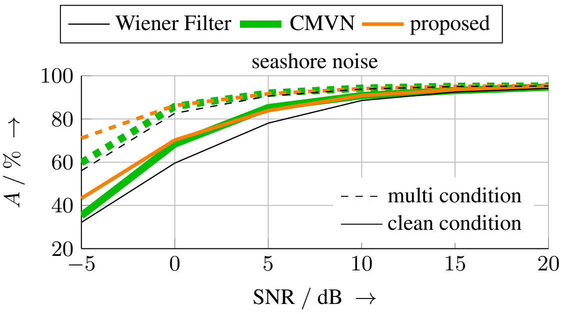

**Fig. 2**: Recognition accuracy measured for various feature processing strategies in seashore noise in dependence on the SNR ($A$: recognition accuracy, SNR: signal-to-noise ratio).

Figure 2 contains the results for the seashore noise. The graphs show that the proposed method increases the speech recognizer's accuracy in comparison to CMVN. This becomes clearly visible in low SNR conditions. At an SNR of $-5$ dB, the accuracy can be boosted by roughly 10 % in absolute value for multi and clean condition training. While the state-of-the-art noise reduction improves the performance when no feature normalization is applied [11], these results show that no improvements are achieved in comparison to CMVN. Here, the obtained accuracy is even slightly lower compared to CMVN especially if the system is trained on clean speech data.

Also for the traffic noise, for which the results are shown in Figure 3, higher recognition accuracies are visible. Again, the largest improvements are obtained in low SNR conditions. At $-5$ dB, the accuracy is 10 % larger in absolute value compared to CMVN if multi condition training is used. For clean condition training, the performance gain is about 5 % in absolute value.

[1] taken from http://www.freesound.org/people/kijjaz/sounds/172601

[2] taken from http://www.freesound.org/people/timsc/sounds/242922

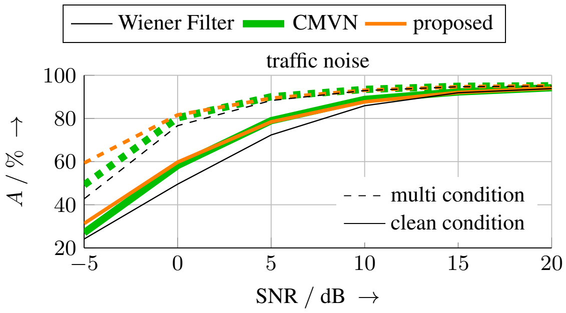

**Fig. 3**: Recognition accuracy measured for various feature preprocessing strategies in traffic noise in dependence on the SNR ($A$: recognition accuracy, SNR: signal-to-noise ratio).

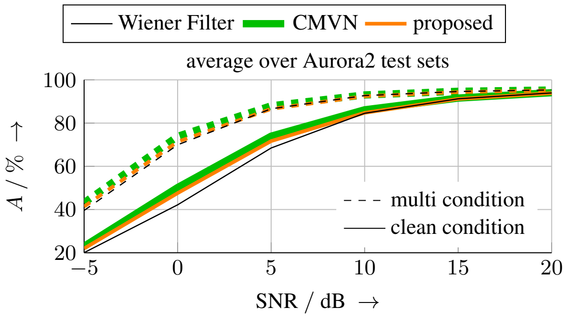

**Fig. 4**: Recognition accuracy measured for various feature preprocessing strategies as mean over all test sets of the Aurora2 database in dependence on the SNR ($A$: recognition accuracy, SNR: signal-to-noise ratio).

In this last paragraph, the results are presented for the three test sets included in the Aurora2 database. The recognition rates are shown in Figure 4. Here, the mean over all noise types of the corpus is shown. In contrast to the other noise types, no improvements are obtained using the proposed method. The accuracy is nearly on the same level as for CMVN. This result can probably be explained by the strongly instationary noise characteristics, e. g. of babble noise, which cannot be tracked precisely by the noise estimator.

## 6. CONCLUSIONS

In this paper, we introduced a novel method for feature enhancement which incorporates techniques used in speech enhancement for improving the performance of ASR. In contrast to methods which enhance the input signal using single-channel noise reduction before the features are extracted, the proposed method is able to increase the recognition rate in combination with CMVN. For this, the MFCC representation of the background noise is subtracted from the noisy MFCCs which effectively reduces the impact of noise modulations onto the MFCC features. If the changes of the background noise are too quick to be captured by CMVN while it is possible to follow these using a state-of-the-art noise estimator, considerable improvements in terms of recognition accuracy are obtained. These emerge especially in low SNR conditions as demonstrated in the experiments conducted for seashore and traffic noise.

## 7. REFERENCES

[1] X. He and L. Deng, “Speech-centric information processing: An optimization-oriented approach,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1116–1135, May 2013.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[3] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[4] O. Viikki and K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” *Speech Communication*, vol. 25, no. 1–3, pp. 133 – 147, 1998.

[5] J. Droppo and A. Acero, “Environmental robustness,” in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer Berlin Heidelberg, 2008, pp. 653 – 679.

[6] J. P. Openshaw and J. Masan, “On the limitations of cepstral features in noise,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Adelaide, South Australia, Australia, Apr. 1994, pp. 49–52.

[7] S. Molau, F. Hilger, and H. Ney, “Feature space normalization in adverse acoustic conditions,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, China, Apr. 2003, pp. 656–659.

[8] T. Gerkmann and R. C. Hendriks, “Noise power estimation based on the probability of speech presence,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, 2011, pp. 145–148.

[9] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[10] C. Breithaupt, T. Gerkmann, and R. Martin, “A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, Apr. 2008, pp. 4897–4900.

[11] C. Breithaupt and R. Martin, “DFT-based speech enhancement for robust automatic speech recognition,” in *ITG Conference on Voice Communication (Sprachkommunikation)*, Aachen, Germany, Oct. 2008.

[12] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge University Engineering Department, Dec. 2006.

[13] H.-G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *ASR-2000 - Automatic Speech Recognition: Challenges for the new Millenium*, Paris, France, Sep. 2000, pp. 181 – 188.