BINAURAL SPEECH ENHANCEMENT WITH INSTANTANEOUS COHERENCE SMOOTHING USING THE CEPSTRAL CORRELATION COEFFICIENT

Rainer Martin, Masoumeh Azarpour, and Gerald Enzner

Institute of Communication Acoustics, Ruhr-Universität Bochum, 44780 Bochum, Germany
{rainer.martin | masoumeh.azarpour | gerald.enzner}@rub.de

ABSTRACT

In this paper we propose a novel approach to cepstral smoothing for reducing musical noise fluctuations in binaural speech enhancement. Similar to other methods, our approach computes a preliminary spectral gain function using the magnitude-squared coherence function and applies an instantaneous weighting to the gain function in the cepstral domain. In this contribution, the weighting function is based on the binaural cepstral correlation coefficient (CCC). We introduce the CCC and briefly discuss its properties. Similar to the cepstrum, the CCC emphasizes the spectral envelope and fundamental frequency information of the target signal, however, in a representation normalized to a range of [-1,1] and less sensitive to spatially uncorrelated noise. Thus, it can be easily and effectively used as a weight in the cepstral domain. The utility of the CCC is confirmed via experiments with different noise types and several instrumental measures.

Index Terms— coherence, cepstrum thresholding, smoothing

1. INTRODUCTION

In this paper we propose a modification of the well-known binaural algorithm [1] for noise and reverberation reduction using the magnitude-squared (MSC) coherence function. This algorithm (denoted as "ABB" in the following) has been originally proposed by Allen, Berkley, and Blauert to suppress reverberation and has been also used in a number of noise reduction applications [2]. In a diffuse noise field or in the reverberant tails of a speech signal the correlation between microphone signals above a certain cut-off frequency is low and can thus be separated from the highly coherent direct sound components. Similar methods, with some variations, were also employed as a postfilter in conjunction with microphone arrays [3], [4]. While averaging a correlation or coherence measure over many microphone pairs of a microphone array leads to a high-quality spectral gain function, the use of this method with only two microphones has always been less satisfying: In this case, the output signal is either plagued by smearing of transient sounds due to the temporal averaging in the estimation of the MSC, or by musical noise if the smoothing is reduced to a minimum.

A new approach to solving the musical noise problem has been proposed in [5, 6] and thoroughly analysed in [7]. The *temporal cepstrum smoothing* approach has been successfully employed to remove musical noise in speech enhancement systems and can be applied to any real-valued gain function. However, it might introduce a slight amount of reverberance. Therefore, in [8] an approach has been introduced which uses *instantaneous cepstrum nulling* (also related to cepstrum thresholding [9], [10]) to remove rapid fluctuations in the spectrum without using a temporal smoothing process. However, this method requires information about the typical range of fundamental frequencies in terms of estimated prior probabilities, and these are acquired in a training step [8].

The proposed, fully adaptive approach uses instantaneous cepstral smoothing and, unlike the method in [8], does not need training data. Towards this end, it is shown that the cepstral correlation coefficient (CCC) is a useful tool. As the normalization of the CCC ensures a range of $[-1, 1]$ it is well suited as a weighting (or soft-thresholding) function in the cepstral domain. In the remainder of this paper we briefly introduce the standard MSC-based method and the CCC. We explain the proposed algorithm and present experimental results in terms of several instrumental measures and the resulting gain functions.

2. COHERENCE-BASED SPEECH ENHANCEMENT

We assume that the microphone signals $y_1(k)$ and $y_2(k)$ are composed of a reverberated target signal $s(k)$ and additive noise $n_{1/2}(k)$

$$y_{1/2}(k) = s(k) * h_{1/2}(k) + n_{1/2}(k) \tag{1}$$

where the indices in the subscript indicate the microphone channels. $h_{1/2}(k)$ denotes the impulse responses between the source and the microphones. Neglecting cyclic effects, these signals can be written in the short-time Fourier domain as

$$Y_{1/2}(m, \mu) = S(m, \mu)H_{1/2}(m, \mu) + N_{1/2}(m, \mu) \tag{2}$$

where m and μ denote the temporal segment (frame) and the frequency bin indices, respectively.

In the next step we compute a preliminary gain function based on the magnitude-squared coherence (MSC) function [1]

$$K(m, \mu) = \frac{|\mathrm{E}\left\{Y_1(m, \mu)Y_2^*(m, \mu)\right\}|^2}{\mathrm{E}\left\{|Y_1(m, \mu)|^2\right\}\mathrm{E}\left\{|Y_2(m, \mu)|^2\right\}}\,. \tag{3}$$

In order to find an estimate $\widehat{K}(m, \mu)$ the statistical expectations in (3) are approximated by first-order recursive systems

$$\begin{aligned}\overline{Y_\kappa(m, \mu)Y_\lambda^*(m, \mu)} &= \alpha\overline{Y_\kappa(m-1, \mu)Y_\lambda^*(m-1, \mu)} \\ &\quad + (1-\alpha)Y_\kappa(m, q)Y_\lambda^*(m, q)\end{aligned} \tag{4}$$

with a smoothing parameter α and $\kappa, \lambda \in \{1, 2\}$. We then apply the cepstral smoothing procedure described in the next section to the estimated gain function $\widehat{K}(m, \mu)$, resulting in a smoothed gain function $K_{ccc}(m, \mu)$. The instantaneously smoothed gain function

Funds of the German Research Foundation (DFG), Collaborative Research Center 823, Subproject B3, are gratefully acknowledged.

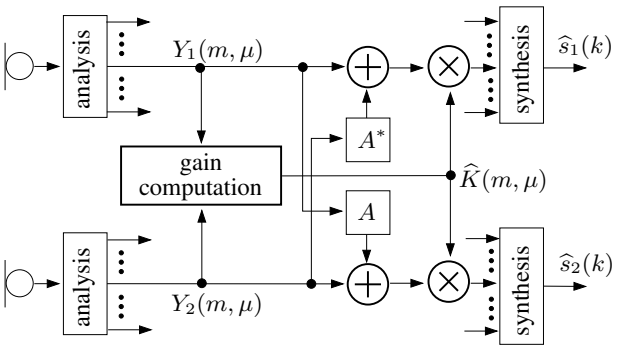

Fig. 1. Binaural noise reduction and dereverberation using the MSC gain function and independent processing in each frequency bin.

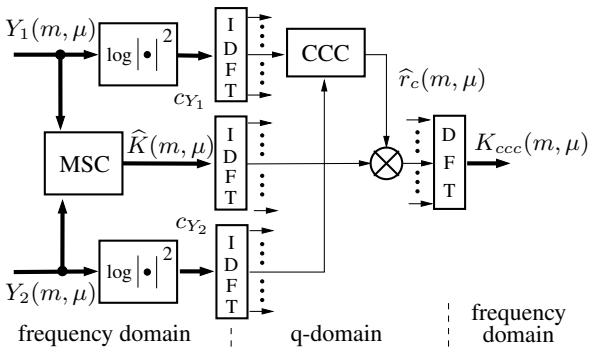

Fig. 2. Instantaneous smoothing using the cepstral correlation coefficient (CCC). Computations in the q-domain are executed independently for each component. Bold arrows signify vectors with elements for $\mu = 0, \ldots, N-1$.

$K_{ccc}(m,\mu)$ and the two input signals are then used to compute binaural output signals. As shown in Fig. 1, we combine the phase-aligned microphone signals in each channel [1]

$$\widehat{S}_1(m,\mu) = \widehat{K}(m,\mu)(Y_1(m,\mu) + A^*(m,\mu)Y_2(m,\mu)), \quad (5)$$

$$\widehat{S}_2(m,\mu) = \widehat{K}(m,\mu)(Y_2(m,\mu) + A(m,\mu)Y_1(m,\mu)), \quad (6)$$

where the phase-alignment function $A(m,\mu)$ is given by

$$A(m,\mu) = \frac{Y_2(m,\mu)Y_1^*(m,\mu)}{\max\left(|Y_1(m,\mu)Y_2(m,\mu)|, \varepsilon\right)} \quad (7)$$

and ε is a small constant in order to prevent divisions by zero. Then, the estimated gain function $\widehat{K}(m,\mu)$ is applied to both channels.

2.1. Temporal Cepstrum Smoothing (CTS)

Smoothing of filter gains in the cepstrum [11] or correlation domains (in either case denoted in what follows by *q-domain*) was proposed in [5] and [6]. In these works, the cepstrum smoothing is implemented in terms of a first-order recursive smoothing system which is applied in each q-domain bin. This smoothing process leads to a significant reduction of musical noise in the resulting smoothed gain $K_{cts}(m,\mu)$ while hardly affecting the quality of the target speech components. However, the temporal smoothing might leave a slightly reverberant effect in the processed signal which is less desirable.

2.2. Instantaneous Cepstral Smoothing

The instantaneous cepstrum smoothing multiplies the q-domain representation (cepstrum or correlation domain) of the gain function for each signal segment with a weighting function. Thereby, speech information is emphasized while noise-dominated bins are attenuated. In this work we propose to derive this weighting function from the correlation coefficient of the cepstrum. As shown in Fig. 2, we compute the cepstral correlation coefficient (CCC)

$$r_c(m,q) = \frac{\mathrm{cov}\{c_{Y_1}(m,q), c_{Y_2}(m,q)\}}{\sqrt{\mathrm{var}\{c_{Y_1}(m,q)\}\mathrm{var}\{c_{Y_2}(m,q)\}}} \quad (8)$$

of the two input signals in each cepstral bin q. The real cepstra [11] for the m-th time frame and the q-th cepstral bin are given by ($\kappa \in \{1,2\}$)

$$c_{Y_\kappa}(m,q) = \frac{1}{N}\sum_{\mu=0}^{N-1} \log\left(|Y_\kappa(m,\mu)|^2\right) \exp\left(j\frac{2\pi\mu q}{N}\right). \quad (9)$$

After the subtraction of the means, the second-order moments are estimated via first-order recursive systems with parameter α_{cc} and $\kappa, \lambda \in \{1,2\}$

$$\begin{aligned}\overline{c}_{Y_\kappa Y_\lambda}(m,q) &= \alpha_{cc}\overline{c}_{Y_\kappa Y_\lambda}(m-1,q) \\ &\quad + (1-\alpha_{cc})c_{Y_\kappa}(m,q)c_{Y_\lambda}(m,q)\,. \end{aligned} \quad (10)$$

The CCC $r_c(m,q)$ may then be approximated by

$$\widehat{r}_c(m,q) = \frac{\overline{c}_{Y_1Y_2}(m,q)}{\sqrt{\overline{c}_{Y_1Y_1}(m,q)\overline{c}_{Y_2Y_2}(m,q)}}\,. \quad (11)$$

The instantaneous smoothing step transforms the preliminary gain function $\widehat{K}(m,\mu)$ in the correlation domain

$$c_K(m,q) = \frac{1}{N}\sum_{\mu=0}^{N-1} \widehat{K}(m,\mu)\exp\left(j\frac{2\pi\mu q}{N}\right). \quad (12)$$

Here, the log-compression of the cepstrum is not necessary [6] as the gain function is restricted to the dynamic of range $[0,1]$. A multiplication of the transformed preliminary gain $c_K(m,q)$ with the CCC weighting function $\widehat{r}_c(m,q)$, i.e., $c_{KS}(m,q) = c_K(m,q)\widehat{r}_c(m,q)$, a subsequent Fourier transform, and a limitation to non-negative values yields the final, smoothed gain

$$K_{ccc}(m,\mu) = \max\left(\sum_{q=0}^{N-1} c_{KS}(m,q)\exp\left(-j\frac{2\pi\mu q}{N}\right), 0\right). \quad (13)$$

Figure 3 shows examples of the gain function before and after applying the smoothing operation. Clearly, during speech pause, undesirable fluctuations are suppressed. Interestingly, also the gain at low frequencies, which is due to the high coherence of the diffuse noise field is reduced. During speech activity, the envelope is well preserved while the dynamic range of the harmonics is only slightly reduced. In Fig. 4 we compare the gain functions $K(m,\mu)$, $K_{cts}(m,\mu)$, and $K_{ccc}(m,\mu)$ for a longer speech sample. Obviously, the instantaneous smoothing preserves the time-frequency structure of the speech signal well while it reduces the spurious fluctuations of gain $K(m,\mu)$ during speech pause.

3. PROPERTIES OF THE CCC

As discussed in [7], the covariance of two cepstral coefficients derived from the same signal is obtained via a 2D Fourier transform

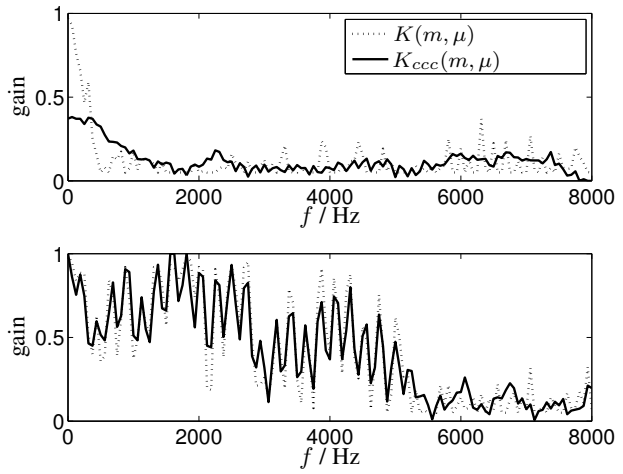

Fig. 3. Gain functions $K(m,\mu)$ and $K_{ccc}(m,\mu)$ for a typical signal segment during speech pause (top) and during voiced speech activity (bottom).

of the covariance of the corresponding log-periodograms. Similarly, we obtain for the covariance of two different cepstra

$$\operatorname{cov}\{c_{Y_1}(m,q), c_{Y_2}(m,q)\} = \frac{1}{N^2}\sum_{\mu_1=0}^{N-1}\sum_{\mu_2=0}^{N-1} e^{\frac{j2\pi q}{N}(\mu_1-\mu_2)} \times \operatorname{cov}\left\{\log\left(|Y_1(m,\mu_1)|^2\right), \log\left(|Y_2(m,\mu_2)|^2\right)\right\} \quad (14)$$

and the same Fourier-transform dependency holds also for the cross-correlation $\mathrm{E}\{c_{Y_1}(m,q)c_{Y_2}(m,q)\}$.

3.1. Uncorrelated Input Signals

To investigate the CCC during speech pause we consider two uncorrelated zero-mean noise signals $y_1(k)$ and $y_2(k)$ and their spectral correlation $\rho(m,\mu) = \mathrm{E}\{Y_1(m,\mu)Y_2(m,\mu)\}$. Since $\rho(m,\mu)$ is zero, the covariance of the log-periodograms and of the cepstral coefficients is zero as well [7]. The cross-correlation of the log-periodograms equals the product of their means

$$\mathrm{E}\left\{\log(|Y_1(m,q)|^2)\log(|Y_2(m,q)|^2)\right\} \quad (15)$$
$$= \mathrm{E}\left\{\log(|Y_1(m,q)|^2)\right\}\mathrm{E}\left\{\log(|Y_2(m,q)|^2)\right\}\ . \quad (16)$$

When the periodograms $|Y_{1/2}(m,\mu)|^2$ obey a bivariate χ^2-distribution with $2L$ degrees of freedom, we have [7]

$$\mathrm{E}\left\{\log\left(|Y_{1/2}(m,\mu)|^2\right)\right\} \quad (17)$$
$$= \psi(L) - \log(L) + \log(\mathrm{E}\left\{|Y_{1/2}(m,\mu)|^2\right\}), \quad (18)$$

and the means of the cepstra are given by

$$\mathrm{E}\left\{c_{Y_{1/2}}(m,q)\right\} = \frac{1}{N}\sum_{\mu=0}^{N-1}\log\left(\sigma^2_{Y_{1/2}}(m,\mu)\right)e^{\frac{j2\pi q}{N}\mu} - \epsilon_q, \quad (19)$$

where $\sigma^2_{Y_{1/2}}(m,\mu) = \mathrm{E}\left\{|Y_{1/2}(m,\mu)|^2\right\}$ and

$$\epsilon_q = \begin{cases}\log(L) - \psi(L) & q = 0\\ 0 & \text{otherwise}\ .\end{cases} \quad (20)$$

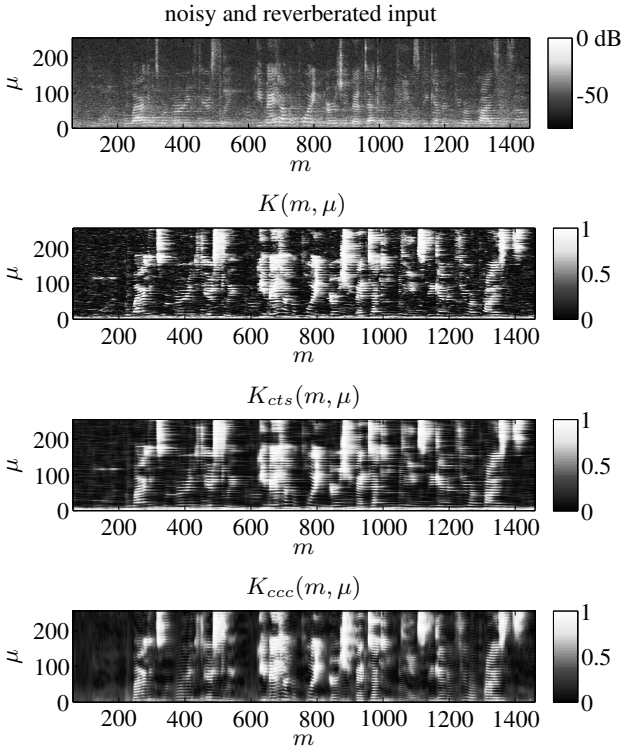

Fig. 4. Spectrograms of reverberant and noisy speech (babble noise) and gain functions $K(m,\mu)$, $K_{cts}(m,\mu)$, and $K_{ccc}(m,\mu)$. The sampling rate is $f_s = 16$ kHz, the frame shift is 5 ms, and the DFT length $N = 512$.

$\psi(L)$ is the psi-function [12, (8.360)]. For speech signals we typically have $0 < L < 1$ while for a Gaussian signal $L = 1$ holds. Hence, for two uncorrelated and spectrally white input signals, the cepstral correlation evaluates to

$$\mathrm{E}\{c_{Y_1}(m,q)c_{Y_2}(m,q)\} = \quad (21)$$
$$\begin{cases}\mathrm{E}\left\{\log\left(|Y_1(m,\mu_1)|^2\right)\right\}\mathrm{E}\left\{\log\left(|Y_2(m,\mu_2)|^2\right)\right\} & q = 0\\ 0 & q \neq 0\end{cases}.$$

The cepstral correlation coefficient in (8) will then have a value close to one for $q = 0$ and smaller magnitudes for $q \neq 0$.

3.2. Harmonic Input Signals

For vowels, the covariance of the dual-channel cepstra shows distinct peaks at the fundamental frequency $F_0 = \mu_0 f_s/N$ and its rahmonics (the term rahmonics is introduced in [11]). The quefrency q_0 of the fundamental frequency and its multiples ℓq_0 are easily derived from (14) with $\mu_0 q_0 = \ell N$ for $\ell = 1, 2, 3, \ldots$. Expanding by f_s/N results in $q_0 = f_s/F_0$ for $\ell = 1$. Thus, similar to the envelope information, the harmonics which both signals have in common are mapped into a few quefrency bins and show up in the cross-channel correlation. The normalization in the denominator of (11) then limits the range to $-1 \leq r_c(m,q) \leq 1$. Finally, due to the instantaneous cepstrum smoothing we may use less smoothing in (4). As a consequence transient speech sounds are less distorted.

3.3. Algorithmic Refinements

Bias Correction: Reduced temporal smoothing of the preliminary MSC estimate $\widehat{K}$ causes a larger bias of the resulting MSC. For a moving average of M independent segments, the expected estimated coherence $E\{\widehat{K}\}$ as a function of the true coherence K is [13]

$$E\{\widehat{K}\} = K + \frac{1}{M}(1-K)^2\left(1+\frac{2K}{M}\right) . \qquad (22)$$

For the recursive smoothing in (4) the effective M is a function of α, the spectral analysis window, and the frame shift [14] but can in general be used to balance noise reduction and speech distortion. Although (22) requires the use of an expected value $E\{\widehat{K}\}$ the approximation $\widehat{K} \approx E\{\widehat{K}\}$ is also effective. Then, (22) can be solved for the true K via a fixed-point iteration with 2-5 steps [15].

Estimation of the Mean Cepstrum: The computation of the CCC requires the subtraction of the mean. However, the estimation of the mean is not very accurate during speech activity. Therefore we estimate the mean only during speech pauses and use the MSC as an indicator of these. As a result we achieve strong interference reduction during speech pauses but also some disturbing discontinuities between speech pause and speech activity. Therefore we chose to subtract the mean only for cepstral bins above a cutoff quefrency.

Thresholding the Cepstral Correlation Coefficient: In general, positive CCC values indicate a synchronous temporal evolution of the two microphone signals and are thus related to the activity of the desired signal. By the same token, CCC values near -1 are not useful for our purposes and should be suppressed. We therefore limit the CCC to values larger than -0.2 and then take the absolute value to smooth out its trajectory.

4. EXPERIMENTAL RESULTS

The performance of proposed algorithm (ABB-CCC) has been comprehensively evaluated and compared to two dual-channel and binaural noise reduction and dereverberation algorithms: the algorithm in [1] (ABB) and the ABB algorithm with a temporal cepstrum smoothing post-processing step (ABB-CTS) [5], [6]. The same spectral, low-delay (10 ms) analysis-synthesis framework [17] has been used for all algorithms. The smoothing parameters have been set to $\alpha = 0.9$ and $\alpha_{cc} = 0.92$. The speech signal is a 60 seconds concatenation of female and male speech taken from TIMIT database [18]. The convolution of clean speech signal with the measured binaural impulse response in lecture room ($T_{60} = 210$ ms) and meeting room ($T_{60} = 700$ ms) from Aachen room impulse response data base [19] are contaminated with different noise types at different levels of SNR. The additive observation noise signals considered in this work are: spherically isotropic and diffuse babble noise which has been generated using the algorithm proposed in [20]. We computed four different performance measures: the segmental SNR improvement (ΔSegSNR), the noise attenuation (NA), the cepstral distance computed on the clean speech signal (CD(s)) [21], and the STOI intelligibility measure [16]. The evaluation results in Fig. 5 show that the ABB-CCC method provides a segmental SNR improvement of up to 2.5 dB while it surpasses the reference methods by almost 1 dB. The highest gains are achieved for input SNR values around 0 dB. The noise attenuation and the cepstral distance measures show that there is a significant improvement in the noise reduction while the distortion of the speech signal does is not increased. This is also evident from the STOI measure.

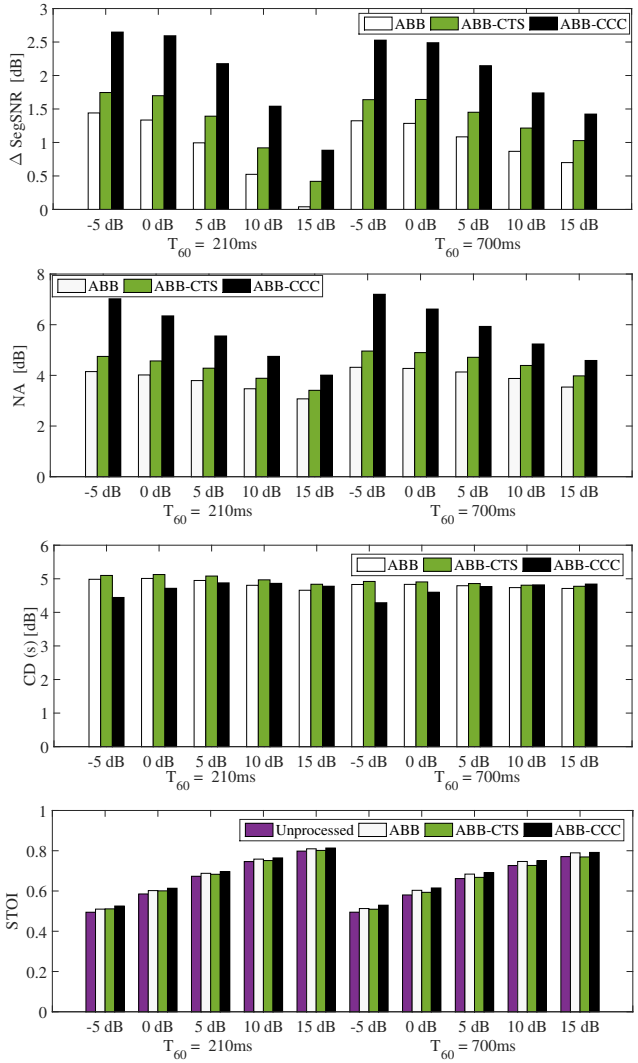

Fig. 5. Experimental results for two rooms and five signal-to-noise ratios {-10, -5, 0, 5, 10}, averaged over two noise types in terms of segmental SNR improvement, noise attenuation, cepstral distance of target speech, and STOI [16]. The smoothing parameters have been set to $\alpha = 0.9$ and $\alpha_{cc} = 0.92$. M in (22) was set to $M = 12$.

5. CONCLUSIONS

This paper introduces the cepstral correlation coefficient (CCC) and explores its use in the context of binaural noise and reverberation reduction. It is shown, that the CCC can be used as a weighting (soft-thresholding) function for cepstral smoothing. This novel smoothing results in a very natural residual noise, and, in contrast to using just the magnitude-squared coherence function to less low-frequency residual noise and musical tones. The CCC is computed online and does not require a training step or the estimation of the fundamental frequency. The experimental results indicate a significant increase in noise reduction with no additional distortions of speech components. The theoretical analysis of the properties of the CCC supports these findings. The CCC weighting emphasizes the salient characteristics of the target speech signal while it smoothes random fluctuations.

6. REFERENCES

[1] J.B. Allen, D.A. Berkley, and J. Blauert, "Multimicrophone Signal-Processing Technique to Remove Room Reverberation from Speech Signals," *J. Acoust. Soc. Am.*, vol. 62, no. 4, pp. 912–915, 1977.

[2] R. Le Bouquin and G. Faucon, "Using the Coherence Function for Noise Reduction," *IEE Proceedings-I*, vol. 139, no. 3, pp. 276–280, 1992.

[3] R. Zelinski, "A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1988, pp. 2578–2581.

[4] I.A. McCowan and H. Bourlard, "Microphone Array Post-filter based on Noise Field Coherence," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, 11 2003.

[5] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral Smoothing of Spectral Filter Gains for Speech Enhancement without Musical Noise," *IEEE Signal Proc. Letters*, vol. 14, no. 12, pp. 1036–1039, 2007.

[6] T. Gerkmann, C. Breithaupt, and R. Martin, "Bias Compensation for Cepstro-Temporal Smoothing of Spectral Filter Gains," in *8. VDE/ITG Conference Speech Communication*, 2008.

[7] T. Gerkmann and R. Martin, "On the Statistics of Spectral Amplitudes After Variance Reduction by Temporal Cepstrum Smoothing and Cepstral Nulling," *IEEE Trans. Signal Processing*, vol. 57, no. 11, pp. 4165–4174, 2009.

[8] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *Proc. Euro. Signal Processing Conf. (EUSIPCO)*, 2011, pp. 2309–2313.

[9] P. Stoica and N. Sandgren, "Smoothed Nonparametric Spectral Estimation via Cepstrum Thresholding," *IEEE Signal Processing Magazine*, vol. 23, pp. 34 – 45, 2006.

[10] P. Stoica and N. Sandgren, "Total-variance Reduction via Thresholding: Application to Cepstral Analysis," *55*, 2007.

[11] B.P. Bogert, M.J.R. Healy, and J.W. Tukey, "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking," in *Proc. of the Symposium on Time Series Analysis*, 1963, pp. 209–243.

[12] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, 6th edition, 2000.

[13] G.C. Carter, "Coherence and time delay estimation," *Proc. of the IEEE*, vol. 75, no. 2, pp. 236–255, Feb 1987.

[14] R. Martin, "Bias Compensation Methods for Minimum Statistics Noise Power Spectral Density Estimation," *Signal Processing, Elsevier*, vol. 86, no. 6, pp. 1215–1229, 2006.

[15] G. Enzner, R. Martin, and P. Vary, "Partitioned Residual Echo Estimation for Frequency Domain Acoustic Echo Control," *European Trans. Telecommunications*, vol. 13, no. 2, pp. 103–114, 2002.

[16] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *TASL*, vol. 19, no. 7, pp. 2125 – 2136, 2011.

[17] D. Mauler and R. Martin, "A Low Delay, Variable Resolution, Perfect Reconstruction Spectral Analysis-Synthesis System for Speech Enhancement," in *Proc. Euro. Signal Processing Conf. (EUSIPCO)*, 2007.

[18] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgrena, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, 1993, ISBN: 1-58563-019-5.

[19] M. Jeub, M. Schäfer, and P. Vary, "A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms," in *Proc. of International Conference on Digital Signal Processing (DSP)*, Santorini, Greece, 2009, pp. 1 – 4.

[20] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating Non-stationary Multisensor Signals Under a Spatial Coherence Constraint," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, 2008.

[21] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Boca Raton, 1st edition, 2007.