TEMPORAL SMOOTHING OF SPECTRAL MASKS IN THE CEPSTRAL DOMAIN FOR SPEECH SEPARATION

Nilesh Madhu, Colin Breithaupt, and Rainer Martin

Institute of Communication Acoustics (IKA)
Ruhr-Universität Bochum, 44780 Bochum, Germany
{firstname.lastname}@rub.de

ABSTRACT

This contribution details the development of a mask-based post- processor to improve the interference suppression in speech signals separated using linear deconvolution algorithms like Independent Component Analysis (ICA). The design of the proposed post-filter is in two stages: in the first stage, use is made of the disjointness of the separated signals in the time-frequency domain to obtain binary masks to suppress cross-talk that generally remains after separation. In the next stage, a novel smoothing of the masks is proposed that preserves the speech structure of the target source while eliminating the random peaks in the time-frequency plane that lead to fluctuating background noise. The result is an enhanced signal with reduced cross-talk and no musical noise.

Index Terms— independent component analysis, time-frequency masking, post processing, cepstro-temporal smoothing, musical noise

1. INTRODUCTION

Source separation techniques are concerned with extracting individual sources from mixtures of competing sources. Generally, these approaches are 'blind' in that no *a priori* knowledge is available regarding the sources. The techniques may be broadly classified into two major categories – linear and non-linear.

Linear deconvolution algorithms for the blind speech separation problem are based on a linear generative mixing model: the observed mixtures are assumed to be linear combinations of the individual speech signals. Consequently, separation is accomplished by convolving the mixtures with the inverse room impulse responses and linearly combining them to yield signals in which only one speaker signal is predominant. Such separation approaches are dependent on the long time statistical properties inherent to speech, e.g., statistical independence between two speaker signals, non-whiteness and non-Gaussianity. These properties are exploited either using higher order statistics (as in the ICA based approaches of, e.g., [1, 2, 3]) or the second order statistics over different time-lags (as in the approaches of [4, 5], to mention a few).

Non-linear approaches to speech signal separation seek an optimal partitioning of the time-frequency (T-F) plane – defined by the discrete Fourier transforms of overlapped, short, windowed signal frames – into T-F components for each individual source. As demonstrated in [6, 7, 8], speech signals of different speakers have mostly non-overlapping supports in the time-frequency plane. This property may be made use of, for partitioning, if one has some *a priori* knowledge regarding which source occupies which T-F point. In such cases, and when the sources are completely disjoint to one another in the T-F plane, one can use binary masks to completely suppress the interfering sources and recover only the desired source. Such mask-based approaches and variations/improvements thereof are detailed in [6, 9, 10, 11, 12].

Obviously, the performance of the mask-based algorithms is contingent upon the validity of the disjoint-support model which, however, does not hold as strongly in reverberant environments. Consequently, these approaches steadily deteriorate in performance as the reverberation increases [8]. Moreover, the masks generated over the T-F plane are very dynamic and often vary significantly from one time-frame to the next. Errors in the estimation of the masks give rise to random isolated peaks in the masked spectrum, resulting in sinusoidal artefacts of one-frame duration and varying frequencies in the re-synthesized signal, which are perceived as so-called *musical noise*.

On the other hand, as linear algorithms are based on long term statistics, the demixing filters, once estimated, remain constant over a relatively long period of time – leading to no musical noise in the separated signals. Further, such approaches perform well even in rather reverberant environments. However, due to the dependence on long term statistics, linear algorithms can only suppress the interference on an *average*. Consequently, interference suppression is obtained to a lesser extent as compared to the mask-based methods.

Performance of the linear algorithms may be enhanced by the use of *post processors* which operate on the signals output by the demixing algorithms. This contribution details the development of such a post processor. The document is structured as follows: Section 2 introduces the signal model and the notations used subsequently in the text. Section 3 describes existing approaches to post-processing followed by their caveats. The proposed approach is then described Section 5 and evaluated via listening tests in the following section.

2. SIGNAL MODEL

Blind source separation (BSS) approaches usually consider a scenario in which Q simultaneously active sources s_q are recorded at M closely spaced microphones in a room:

$$\begin{bmatrix} \tilde{x}_1(t) \\ \vdots \\ \tilde{x}_M(t) \end{bmatrix} = \begin{bmatrix} \tilde{h}_{11}(t) & \cdots & \tilde{h}_{1Q}(t) \\ \vdots & \ddots & \vdots \\ \tilde{h}_{M1}(t) & \cdots & \tilde{h}_{MQ}(t) \end{bmatrix} * \begin{bmatrix} \tilde{s}_1(t) \\ \vdots \\ \tilde{s}_Q(t) \end{bmatrix}, \quad (1)$$

where the $\tilde{h}_{mq}(t)$ represent the room-impulse responses from source q to microphone m and $*$ represents the convolution operator. Usually separation is done in the short-time discrete Fourier domain,

The work of C. Breithaupt is funded by the German Research Foundation DFG.

where the mixing model may be approximated as:

$$\begin{bmatrix} x_1(k,n) \\ \vdots \\ x_M(k,n) \end{bmatrix} = \begin{bmatrix} h_{11}(k) & \cdots & h_{1Q}(k) \\ \vdots & \ddots & \vdots \\ h_{M1}(k) & \cdots & h_{MQ}(k) \end{bmatrix} \begin{bmatrix} s_1(k,n) \\ \vdots \\ s_Q(k,n) \end{bmatrix}, \quad (2)$$

where k represents the frequency bin index and n the time-frame under consideration and the convolution is replaced by the multiplication operator. Note that the spectral model in (2) implies an approximation of the $\tilde{h}_{mq}$ by finite impulse responses.

Linear separation algorithms then seek optimal filters $w_{qm}(k)$ such that the de-mixed signals $y_q(k,n)$ predominantly contain only one source signal as received at the microphone, i.e.,

$$\begin{aligned} \begin{bmatrix} y_1(k,n) \\ \vdots \\ y_Q(k,n) \end{bmatrix} &= \begin{bmatrix} w_{11}(k) & \cdots & w_{1M}(k) \\ \vdots & \ddots & \vdots \\ w_{Q1}(k) & \cdots & w_{QM}(k) \end{bmatrix} \begin{bmatrix} x_1(k,n) \\ \vdots \\ x_M(k,n) \end{bmatrix} \\ &= \begin{bmatrix} a_{11}(k) & \cdots & a_{1Q}(k) \\ \vdots & \ddots & \vdots \\ a_{Q1}(k) & \cdots & a_{QQ}(k) \end{bmatrix} \begin{bmatrix} s_1(k,n) \\ \vdots \\ s_Q(k,n) \end{bmatrix} \end{aligned} \quad (3)$$

where $a_{qq}(k) \approx h_{qq}(k)$ to solve the scaling uncertainty [2, 13] and $a_{qm}(k) \approx 0,\ m \neq q$, for interference suppression.

Usually the anti-diagonal terms of the filter matrix in equation (3) are not zero due to the approximation in (2), errors in the estimation of the demixing filters and due to the approximation of the room impulse response by finite-length FIR filters, giving rise to cross-talk, which is especially disturbing during the speech pauses of the target speaker.

3. POST-PROCESSING

To obtain enhanced separation in reverberant environments, the reverberation-robust linear algorithms in (3) are first used to obtain an approximation to the separated signals ($y_q(k,n) \approx h_{qq}(k)s_q(k,n)$). Next, assuming the disjointness [7] of the underlying, true, source signals $s_q(k,n)$, we have

$$s_q(k,n)s_{q'}(k,n) = 0,\ \forall q' \neq q. \quad (4)$$

In other words, only one speaker is *dominant* at any one T-F point (k,n). Consequently, we may conclude, for instance, when any one recovered signal has more energy than the others at any T-F point, the corresponding source is dominant for that point and appears in the other signals as interference. Thus we may define suitable masks $\mathcal{M}_q(k,n)$ in the time frequency domain to block out such cross-talk. In their simplest form these masks are defined as:

$$\mathcal{M}_q(k,n) = \begin{cases} 1 & \gamma\,|y_q(k,n)| > \max\limits_{\forall q' \neq q} (|y_{q'}(k,n)|) \\ \mathcal{M}_{\text{min}} & \text{else} \end{cases} \quad (5)$$

where $0 < \gamma \leq 1$ is used to prevent spurious triggering of the masks, and where $\mathcal{M}_{\text{min}}$ is the maximum suppression allowed. The final, enhanced signals are subsequently obtained as

$$z_q(k,n) = \mathcal{M}_q(k,n)y_q(k,n), \quad (6)$$

from which the discrete time signals may be reconstructed by the inverse discrete Fourier transform, followed by standard overlap-add procedures.

Such post-processors based on binary masks have been proposed in, e.g., [14] (for various γ). Further improvements to mask-based post-filters include that presented in [12] and [15].

4. CAVEATS OF POST-PROCESSING

While the mentioned post-processing approaches enhance the suppression of the interference signal, the time-variant nature of the masks in the T-F plane introduces musical noise into the signal. One way to avoid this harmonic distortion is to smooth the masks along time and/or frequency. However, simple temporal smoothing delays the response of the masks to speech onsets and smoothing along frequency has the effect of reducing the spectral resolution – smearing the signal across the spectrum. Another way is to use *soft* masks, which alleviate this problem by *limiting* the achievable suppression. However, the drawback here is that the interfering signal is suppressed to a lower extent as compared to when binary masks are used.

Therefore, for good interference suppression and no musical noise we shall use binary masks followed by an optimal smoothing algorithm that is able to distinguish between unwanted isolated random peaks in the mask $\mathcal{M}_q(k,n)$ on one side and mask patterns resulting from the spectral structures of the target speech on the other. Such a smoothing is the topic of the following section.

5. MASK SMOOTHING IN THE CEPSTRAL DOMAIN

As speech signals, in general, have a broad-band envelope, a temporal smoothing should not be applied to the mask when the broad-band structure of the signal changes. Likewise, a change in the fine structure of the spectrum that originates from an onset of voiced speech (pitch harmonics) must also be protected from smoothing effects. Ideally, the smoothing should only affect irregular peaks of short duration. This distinction between the speech related broad-band structures and regular pitch harmonics on one side and the irregular fine-structured artefacts like isolated random peaks on the other is accomplished in the *cepstral* [16] domain. Consequently, the cepstral representation of the mask pattern – $\mathcal{M}_q^{\text{cepst}}(l,n)$ – is first obtained as:

$$\mathcal{M}_q^{\text{cepst}}(l,n) = \text{DFT}^{-1}\big\{ \ln(\mathcal{M}_q(k,n))\big|_{k,=0,\ldots,K-1} \big\}, \quad (7)$$

where l is the quefrency bin index, $\text{DFT}\{\,\cdot\,\}$ represents the discrete Fourier transform operator, and K is the length of the transform. Next a first order, temporal, recursive smoothing is applied to $\mathcal{M}_q^{\text{cepst}}(l,n)$ as:

$$\overline{\mathcal{M}}_q^{\text{cepst}}(l,n) = \beta_l\, \overline{\mathcal{M}}_q^{\text{cepst}}(l,n-1) + (1-\beta_l)\, \mathcal{M}_q^{\text{cepst}}(l,n)\,, \quad (8)$$

where the smoothing constants β_l are chosen separately for the different quefrency bins l according to:

$$\beta_l = \begin{cases} \beta_{\text{env}} & \text{if } l \in \{0,\ldots,l_{\text{env}}\}\,, \\ \beta_{\text{pitch}} & \text{if } l = l_{\text{pitch}}\,, \\ \beta_{\text{peak}} & \text{if } l \in \{(l_{\text{env}}+1),\ldots,K/2\} \setminus \{l_{\text{pitch}}\}\,. \end{cases} \quad (9)$$

The rationale behind this choice for β_l is as follows: for the lower bins $l \in \{0,\ldots,l_{\text{env}}\}$, the values of $\mathcal{M}_q^{\text{cepst}}(l,n)$ represent the *spectral* envelope of the mask $\mathcal{M}_q(k,n)$ [16]. As speech onsets go along with a sudden rise in the spectral envelope, β_{env} should have a very low value, resulting in a low smoothing, in order not to distort the envelope. Likewise, if l_{pitch} is the quefrency bin that represents the regular structure of the pitch harmonics in $\mathcal{M}_q(k,n)$ [16], we apply a relatively low smoothing β_{pitch} to this bin ($l = l_{\text{pitch}}$). The cepstral bins $l \in \{(l_{\text{env}}+1),\ldots,K/2\} \setminus \{l_{\text{pitch}}\}$ represent the fine structure of $\mathcal{M}_q(k,n)$ that is not related to the pitch and cover, with high

probability, the random unwanted peaks that lead to the harmonic distortion. Therefore, we apply a strong smoothing $\beta_{\text{peak}} (> \beta_{\text{pitch}})$ to these coefficients. As the unwanted isolated random peaks represent a sporadic change of the fine structure of $\mathcal{M}_q(k,n)$, and as they last only for a short duration, they are strongly affected by the smoothing (8). Note that this smoothing does not affect the speech information contained in the upper quefrency bins (apart from l_{pitch}) as such information is generally present for more than one frame and are thus preserved despite the high value of β_{peak}.

For the frame n under consideration, we choose l_{pitch} as the cepstral bin that satisfies

$$l_{\text{pitch}} = \underset{l}{\text{argmax}} \left\{ \mathcal{M}_q^{\text{cepst}}(l,n) \big| l_{\text{low}} \leq l \leq l_{\text{high}} \right\}, \qquad (10)$$

which is a well-known method for computing the pitch frequency from a cepstrum [16]. The search range $\{l_{\text{low}}, l_{\text{high}}\}$ is selected so that possible pitch frequencies between 70 Hz and 500 Hz may be detected. Although the search in equation (10) only delivers meaningful results in the presence of voiced speech, the signal energy contained in the bin l_{pitch}, otherwise, is comparably low so that no audible side effects are perceivable from the lesser smoothing $\beta_{\text{pitch}} < \beta_{\text{peak}}$ of that bin in the absence of voiced speech.

For bins $l > K/2$, $\overline{\mathcal{M}}_q^{\text{cepst}}(l,n)$ is determined by the symmetry condition of the DFT: $\overline{\mathcal{M}}_q^{\text{cepst}}(l,n) = \overline{\mathcal{M}}_q^{\text{cepst}}(K-l,n)$. The final smoothed *spectral* mask is obtained as:

$$\overline{\mathcal{M}}_q(k,n) = \exp\left(\text{DFT}\{\overline{\mathcal{M}}_q^{\text{cepst}}(l,n)\big|_{l=0,\ldots,K-1}\}\right), \qquad (11)$$

where the exponential function is applied element-wise. This smoothed mask is then used to obtain the enhanced signal according to (6).

6. EVALUATION & DISCUSSION

The test data used to evaluate the proposed post-processor consists of the individual separated signals from a $Q = 2$ source, $M = 2$ microphone, ICA based algorithm (according to [3]) operating on mixtures recorded in an office room with a reverberation time $T_{60} = 0.5$s. The time-domain signals were segmented into frames of length K and weighted by a Hann window before transformation into the discrete Fourier domain. The overlap between adjacent frames was set to 50%. Table 1 summarizes the values of the remaining parameters for the system used in the evaluation.

$f_s = 8\,\text{kHz}$	$l_{\text{env}} = 8$	$\beta_{\text{env}} = 0$
$K = 256$	$l_{\text{low}} = 16$	$\beta_{\text{pitch}} = 0.4$
$20\log_{10}(\gamma) = -5\text{dB}$	$l_{\text{high}} = 120$	$\beta_{\text{peak}} = 0.8$
$\mathcal{M}_{\text{min}} = 0.1$		

Table 1. Parameter values for the implemented post-processing system.

The proposed post-processing algorithm was evaluated via listening tests on eleven test-subjects, encompassing both 'expert' and 'non-expert' listeners. The test set consisted of 24 examples containing mixtures of male-male, male-female and female-female speakers. For each example, the test subject was presented three audio samples – the output of the ICA algorithm (without post processing), the output after post processing the signals obtained from the ICA by the binary masks as in equation (5) and, finally, the output of the proposed post-processor (the masks of (5) smoothed as proposed in Section 5). The test subject was then asked to select the best of the three for each category:

1. quality of speech of the desired speaker,
2. interference suppression, and
3. overall impression.

The purpose of the listening test was to first confirm that the post-processing using binary masks (5) improved interference suppression and, secondly, to verify that the proposed method improved the masks by removing disturbing musical noise without degrading the high interference suppression, resulting in a better, overall acceptance. The results are presented in Figure 1.

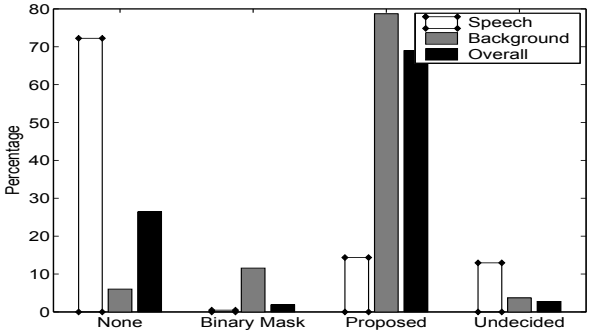


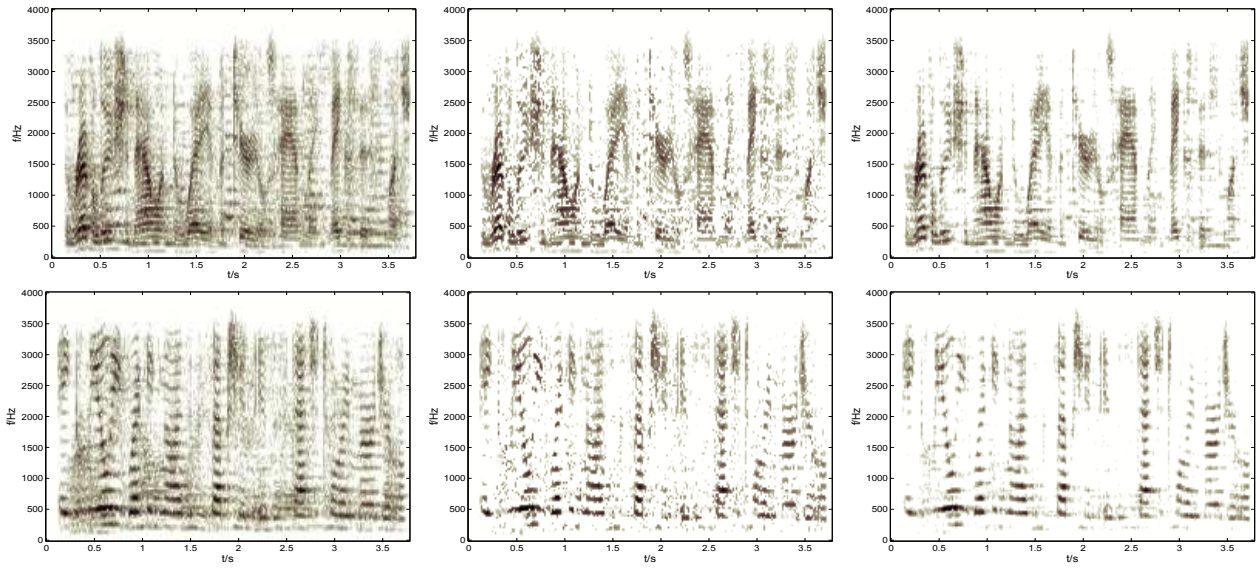
Fig. 1. Cumulative results of the listening test in terms of speech quality (Speech), interference suppression (Background) and overall impression ('none' indicates no *post*-processing).

Additionally, Figure 2 shows the spectra of the signals recovered by the ICA, by the simple binary masks and by the proposed post-filter. Note the profusion of isolated peaks in Figure 2 (b, top and bottom) for the binary mask, and the corresponding version in Figure 2 (c, top and bottom), obtained from the proposed approach – where the peaks are successfully suppressed. Note also that the proposed approach preserves the speech onsets and the pitch structure in the recovered signals.

From the listening test results, it is apparent that a post-processing (5) indeed reduces the amount of interference (Figure 1) – indicating that it is useful to implement a mask-based post processor despite the signals being reverberant. For the evaluation of the quality of speech, most subjects opted for the signal with no post-processing as having the best quality. This is because, due to masking and the threshold γ, there is a slight distortion in the post-processed signal spectrum. Note that this originates from the binary mask of (5) and not from the proposed smoothing and could be dealt with by more sophisticated masking approaches. However, the subjects did state that, at times, they found it hard to distinguish between the speech quality afforded by the proposed post-processor and that of ICA. This is reflected in the relatively large number of 'undecided' votes. In terms of overall impression, the proposed method delivers the best performance – indicating the merit of the approach.

7. CONCLUSIONS

In summary, the proposed post-filter, consisting of a binary mask followed by a first-order recursive, temporal smoothing in the cepstral domain is effective in reducing cross-talk without the unwanted and annoying side-effect of musical noise. Additionally, smoothing in the cepstral domain makes it easier to preserve the spectral characteristics of the target speaker while smoothing out the effects of unwanted, random peaks in the spectrum. Thus, the proposed post-processor provides a high interference suppression and, simultaneously, prevents musical noise. The listening tests corroborate



(a) Output of the ICA algorithm. (b) Spectrum after binary masking. (c) Spectrum after smoothing masks.

Fig. 2. Spectra of the recovered signals before and after post-processing. Each row shows the spectrograms for one speaker.

our conclusions. Note that this method is not restricted to the post-processing for source separation. In general, it lends itself readily as a smoothing approach in cases where the disjointness property of speech is used to compute time and frequency-variant gains for desired speech enhancement. As an example, this approach could also be used directly on mask-based separation approaches, or in noise reduction algorithms [17]. However, for this specific case, we find it best to use the masking approach in conjunction with the linear algorithm, followed by the post filtering proposed here.

8. REFERENCES

[1] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind Source Separation based on a Fast-Convergence algorithm combining ICA and Beamforming," *IEEE TASLP*, vol. 14, pp. 666–678, Mar. 2006.

[2] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE TSAP*, vol. 12, pp. 530–538, Sept. 2004.

[3] N. Madhu, A. Gückel, and R. Martin, "Combined beamforming and frequency domain ICA for source separation," in *Proc. IWAENC*, Sept. 2006.

[4] C. L. Fancourt and L. Parra, "The coherence function in blind source separation of convolutive mixtures of non-stationary signals," in *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, 2001, pp. 303–312.

[5] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," in *Proc. IEEE ICASSP*, 2004.

[6] Ö. Yilmaz, A. Jourjine, and S. Rickard, "Blind separation of disjoint orthogonal signals: Demixing n sources from two mixtures," in *Proc. IEEE ICASSP*, 2000.

[7] S. Rickard and Ö. Yilmaz, "On the approximate W-Disjoint orthogonality of speech," in *Proc. IEEE ICASSP*, 2002.

[8] Ö. Yilmaz, A. Jourjine, and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *Proc. IEEE TSP*, vol. 52, no. 7, July 2004.

[9] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," in *Proc. International Conference on Independent Component Analysis (ICA)*, Dec. 2001.

[10] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. IWAENC*, Sept. 2005.

[11] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proc. IEEE ICASSP*, 2005, pp. 81–84.

[12] F. Flego, S. Araki, H. Sawada, T. Nakatani, and S. Makino, "Underdetermined blind separation for speech. in real environments with F0 adaptive comb filtering," in *Proc. IWAENC*, 2005.

[13] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," in *Proc. International Conference on Independent Component Analysis (ICA)*, Dec. 2001, pp. 722–727.

[14] D. Kolossa and R. Orglmeister, *Nonlinear Postprocessing for Blind Speech Separation*, vol. 3195 of *Lecture Notes in Computer Science*, pp. 832–839, Springer Verlag, Berlin, 2004.

[15] R. Aichner, M. Zourub, H. Buchner, and W. Kellermann, "Residual cross-talk and noise suppression for convolutive blind source separation," in *Proceedings of the German Acoustical Society (DAGA)*, 2006.

[16] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, 1975.

[17] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE SPL*, vol. 14, no. 12, December 2007.