SPEECH PRESENCE PROBABILITY ESTIMATION BASED ON TEMPORAL CEPSTRUM SMOOTHING

Timo Gerkmann, Martin Krawczyk, and Rainer Martin

Institute of Communication Acoustics (IKA)
Ruhr-Universität Bochum, 44780 Bochum, Germany
{timo.gerkmann,martin.krawczyk,rainer.martin}@rub.de

ABSTRACT

We propose a novel, robust estimator for the probability of speech presence at each time-frequency point in the short-time discrete Fourier domain. While existing estimators perform quite reliably in stationary noise environments, they usually exhibit a large false-alarm rate in nonstationary noise that results in a great deal of noise leakage when applied to a speech enhancement task. The proposed estimator overcomes this problem by temporally smoothing the cepstrum of the *a posteriori* signal-to-noise ratio (SNR), and yields considerably less noise leakage and low speech distortions in both, stationary and nonstationary noise as compared to state-of-the-art estimators. Especially in babble noise, this results in large SNR improvements.

Index Terms— Speech presence probability, speech analysis, cepstral analysis, speech enhancement, smoothing methods.

1. INTRODUCTION

In many areas of speech processing an estimate of the probability of speech presence is required. Such an estimate can, for instance, increase the performance of single channel speech enhancement algorithms [1, 2, 3, 4, 5, 6], or can be used in multichannel speech enhancement to discard channels that are more severely disturbed than others [7].

The estimation of speech presence at each time-frequency point in the short-time discrete Fourier domain is a challenging task. Estimators as presented in [2, 3] exhibit only little speech distortion, but do not yield low speech presence probability (SPP) estimates in speech absence. Estimators like [4, 5] overcome this problem, but still exhibit severe noise leakage in nonstationary environments, such as in babble noise. Furthermore, the estimators in [4, 5] require the combination of two SPP estimates based on local and global spectral smoothing.

State-of-the-art estimators for the noise power are often based on the fact that noise is more stationary than speech [8], and are consequently not capable of tracking instationarities, *e.g.* high energy noise bursts of short duration that often occur in babble noise. Hence, noise bursts are misinterpreted as speech and SPP estimators like [2, 3, 4, 5] exhibit a high false-alarm rate in nonstationary noise.

Recently, it has been shown that a temporal smoothing of the cepstral representation of certain spectral quantities performs better than a smoothing in the frequency domain [9]. As in the cepstral domain speech is represented by only few coefficients, a selective smoothing of the speech related coefficients is possible, which enables the elimination of spectral outliers caused by local underestimations of the noise power without affecting the speech signal.

This work combines recent findings on *a posteriori* SPP estimation [5] (Section 2), the idea of cepstral smoothing presented in [9] (Section 3), and the effect of a cepstral smoothing on the statistics of χ^2-distributed random variables presented in [10] (Section 4). As a result, we present a new estimator for the *a posteriori* SPP that clearly outperforms state-of-the-art SPP estimators in nonstationary noise and also achieves better performance in stationary noise (Section 4).

2. A POSTERIORI SPP ESTIMATION

We assume an additive mixture of speech, $S_k(l)$, and noise, $N_k(l)$, in the short-time discrete Fourier domain. Here, k is the frequency index and l is the segment index. The observed signal under the hypothesis $\mathcal{H}_1$, which signifies the presence of speech, is given as $Y_k(l) = S_k(l) + N_k(l)$. Whereas, under hypothesis $\mathcal{H}_0$ that indicates the absence of speech, the observed signal takes the form $Y_k(l) = N_k(l)$. In the following, we whenever possible omit the frame index l for notational convenience. We assume that the spectral noise power $\sigma^2_{\mathrm{N},k} = \mathrm{E}\{|N_k|^2\}$ is available, and introduce the normalized observation $\gamma_k = |Y_k|^2/\sigma^2_{\mathrm{N},k}$ as the *a posteriori* signal-to-noise ratio (SNR). In practice, we estimate the noise power using the minimum statistics approach [8]. For the short-time Fourier analysis we use Hann windows with a length of 32 ms and 50% overlap. The signals are sampled at 16 kHz.

In all papers mentioned above [1, 2, 3, 4, 5, 6, 7], an *a posteriori* SPP estimate is gained as

$$\mathcal{P}_k = P\{\mathcal{H}_1 | \gamma_k\} = \frac{\Lambda_k}{1 + \Lambda_k} \,. \tag{1}$$

The generalized likelihood ratio (GLR), Λ_k, is defined as the weighted ratio of the likelihoods of speech presence and absence:

$$\Lambda_k = \frac{\rho \quad p\left(\gamma_k \mid \mathcal{H}_1\right)}{(1-\rho) \; p\left(\gamma_k \mid \mathcal{H}_0\right)} \,, \tag{2}$$

where $\rho = P\{\mathcal{H}_1\}$ is the *a priori* SPP.

As in [2, 3, 4, 6, 7] it is assumed that Y_k is complex-Gaussian distributed, which results in a χ^2-distribution with two degrees of freedom for the *a posteriori* SNR γ_k. In the following, the degrees of freedom are expressed by the shape parameter $\mu = r/2$, where r denotes the degrees of freedom. Note, that super-Gaussian distributions for Y_k can be accounted for by setting $\mu < 1$ as proposed in [11]. As in [5] we propose to smooth the *a posteriori* SNR γ_k and denote the smoothed quantity by $\overline{\gamma}_k$. The smoothed random variable remains approximately χ^2-distributed [10, 12]. However, as detailed in Section 4, the smoothing process results in an increase

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

in the shape parameter, *i.e.* $\bar{\mu} > \mu$, where $\bar{\mu}$ is the shape parameter after smoothing. As the *a posteriori* SNR is normalized on the noise power, in speech absence we have $\mathrm{E}\{\overline{\gamma}_k\} = 1$ and

$$p\left(\overline{\gamma}_k \mid \mathcal{H}_0\right) = \frac{1}{\Gamma(\bar{\mu})}\,\bar{\mu}^{\bar{\mu}}\,\overline{\gamma}_k^{\bar{\mu}-1}\exp(-\bar{\mu}\,\overline{\gamma}_k) \ . \tag{3}$$

Assuming that speech and noise are uncorrelated, we have in speech presence

$$p\left(\overline{\gamma}_k \mid \mathcal{H}_1\right) = \frac{1}{\Gamma(\bar{\mu})}\left(\frac{\bar{\mu}}{(1+\xi_k)}\right)^{\bar{\mu}}\overline{\gamma}_k^{\bar{\mu}-1}\exp\!\left(-\bar{\mu}\,\frac{\overline{\gamma}_k}{(1+\xi_k)}\right) , \tag{4}$$

where $\xi_k = \sigma^2_{\mathrm{S},k}/\sigma^2_{\mathrm{N},k}$ is the *a priori* SNR and $\sigma^2_{\mathrm{S},k} = \mathrm{E}\big\{|S_k|^2\big\}$. The GLR results in

$$\Lambda(\overline{\gamma}_k) = \frac{\rho}{1-\rho}\cdot\left(\frac{1}{1+\xi_k}\right)^{\bar{\mu}}\exp\!\left(\frac{\xi_k}{1+\xi_k}\,\bar{\mu}\,\overline{\gamma}_k\right), \tag{5}$$

which is then used in (1) to compute the *a posteriori* SPP $\mathcal{P}_k$.

The likelihood ratio indicates speech presence, if $p\left(\overline{\gamma}_k \mid \mathcal{H}_1\right) > p\left(\overline{\gamma}_k \mid \mathcal{H}_0\right)$ and *vice versa*. The likelihoods of speech absence and presence differ only in their mean value $\mathrm{E}\{\overline{\gamma}|\mathcal{H}_0\} = 1$ and $\mathrm{E}\{\overline{\gamma}|\mathcal{H}_1\} = (1+\xi_k)$, respectively. In [2, 3, 4] the *a priori* SNR ξ_k is estimated using the *decision-directed* approach as proposed by Ephraim and Malah [2]. However, as in speech absence the resulting *a priori* SNR estimate is close to zero ($\xi_k \to 0$), the likelihoods of speech presence and speech absence are identical, and the *a posteriori* SPP yields the *a priori* SPP ρ. Thus, in [3, 4] it is proposed to adaptively learn the *a priori* SPP. However, the adaptation of the *a priori* SPP can be seen as only circumventing the true problem: the likelihoods $p\left(\overline{\gamma}_k \mid \mathcal{H}_0\right)$ and $p\left(\overline{\gamma}_k \mid \mathcal{H}_1\right)$ still tend to be equal in the absence of speech, which signifies a discrepancy in the basic probabilistic model.

In [5] we argue that for SPP estimation neither the *a priori* SNR nor the *a priori* SPP should be adapted, but reflect true *a priori* knowledge. In particular, in order to obtain a reasonable SPP estimate, the *a priori* SNR should reflect the SNR that is expected *if speech were present*. We find an optimal, fixed *a priori* SNR that minimizes the total probability of error, as detailed in [5]. An erroneous estimate is given if $\mathcal{P}_k < 0.5$ in the presence of speech (missed-hit rate) and if $\mathcal{P}_k > 0.5$ in the absence of speech (false-alarm rate).

As in [1] we assume that the speech and noise states are equally likely and use the fixed *a priori* SPP $\rho = 0.5$.

3. SMOOTHING THE NOISY OBSERVATION

Usage of γ_k instead of $\overline{\gamma}_k$ in (5), results in a large amount of outliers in the estimate of the SPP $\mathcal{P}_k$. The outliers can cause annoying artifacts if the SPP estimate is applied to a speech enhancement task, and should thus be avoided. The amount of outliers in $\mathcal{P}_k$ can be mitigated by reducing the variance by smoothing γ_k over time and/or frequency. This reduction of variance in turn decreases the overlap of the likelihoods (3) and (4), and thus results in a lower false-alarm and missed-hit rate, as shown in [5].

A drawback of smoothing over time and/or frequency is that the temporal and/or frequency resolution is reduced. Recently, it has been shown that smoothing in the cepstral domain outperforms smoothing in the frequency domain [9], as speech is very compactly represented in the cepstral domain. The speech related cepstral coefficients are given by few lower cepstral coefficients representing the speech spectral envelope and a peak in the upper cepstrum that represents the speech fundamental period. Consequently, a selective smoothing of speech and the remaining coefficients can be carried out. The selective smoothing allows for a strong reduction of spectral outliers with very little speech distortion. Therefore, in this work we propose to selectively smooth the *a posteriori* SNR in the cepstral domain. For this, the *a posteriori* SNR is transformed into the cepstral domain

$$\gamma_q^{\mathrm{ceps}}(l) = \mathrm{IDFT}\{\,\log(\gamma_k(l))\}\ , \tag{6}$$

recursively smoothed over time

$$\overline{\gamma}_q^{\mathrm{ceps}}(l) = \alpha_q(l)\,\overline{\gamma}_q^{\mathrm{ceps}}(l-1) + (1-\alpha_q(l))\,\gamma_q^{\mathrm{ceps}}(l)\ , \tag{7}$$

and transformed back into the frequency domain

$$\widetilde{\gamma}_k(l) = \exp\!\big(\mathrm{DFT}\big\{\overline{\gamma}_q^{\mathrm{ceps}}(l)\big\}\big)\ . \tag{8}$$

In (6), (7), (8) $\mathrm{DFT}\{\cdot\}$ is the discrete Fourier transform while $\mathrm{IDFT}\{\cdot\}$ is its inverse, q is the cepstrum index (often referred to as *quefrency* index), $\log(\cdot)$ is the natural logarithm, and α_q is a quefrency dependent smoothing constant. Due to the nonlinear $\log$ in (6), the unbiased smoothing in the cepstral domain (7) results in a bias in the spectral domain. In this work we denote a biased, smoothed observation as $\widetilde{\gamma}$, while an unbiased smoothed observation is given as $\overline{\gamma}$.

To achieve a large variance reduction without speech distortion, the smoothing constant α_q in (7) is chosen to be close to zero for the speech related cepstral coefficients and close to one for the remaining coefficients.

4. DETERMINATION OF THE SHAPE PARAMETER AND BIAS COMPENSATION

With [13, (3.462.9)] the moments of a χ^2-distributed random variable can be computed, and it can be shown that the mean and variance are related to $\bar{\mu}$ as

$$\bar{\mu} = (\mathrm{E}\{\overline{\gamma}_k\})^2/\mathrm{var}\{\overline{\gamma}_k\}\ . \tag{9}$$

From (9) it can be seen that an unbiased smoothing necessarily results in an increase in the shape parameter of a χ^2-distributed random variable. Then, for a noise-only signal with $\mathrm{E}\{\overline{\gamma}\} = 1$ the shape parameter is simply given by the reciprocal of the reduced variance $\bar{\mu} = 1/\mathrm{var}\{\overline{\gamma}_k\}$.

However, due to the nonlinear $\log$ in (6) the cepstrum smoothing (7) results in a bias in the frequency domain. Then, the shape parameter $\bar{\mu}$ and the relative bias in the mean of the cepstrally smoothed spectral observation can be determined as derived in [10]. There, it is shown that the shape parameter of a χ^2-distributed random variable after smoothing is directly related to the quefrency sum of the cepstral variance after smoothing, as

$$\zeta(2,\bar{\mu}(l)) = \sum_{q=0}^{K/2}\nu_q\mathrm{var}\big\{\gamma_q^{\mathrm{ceps}}\big\}\,\frac{1-\alpha_q(l)}{1+\alpha_q(l)}\,, \tag{10}$$

with

$$\nu_q = \begin{cases} 1/2 & , q \in \{0, K/2\} \\ 2 & , \text{else} \end{cases}\ , \tag{11}$$

$\zeta(\cdot,\cdot)$ Riemann's zeta-function [13, (9.521.1)], and K the length of the Fourier transform used for the short-time analysis. The factor $\frac{1-\alpha_q}{1+\alpha_q}$ follows from the variance reduction in the cepstral domain as achieved by (7). In practice, the relation between $\bar{\mu}$ and $\zeta(2,\bar{\mu})$ can be stored in a table. The variance of cepstral coefficients before cepstral smoothing for $\mu = 1$ is derived to be [10]:

$$\mathrm{var}\{\gamma_q^{\mathrm{ceps}}\} \approx \begin{cases} \frac{2}{K}\left(\frac{\pi^2}{6}+\cos\left(\frac{2\pi}{K}q\right)\right) & ,q\in\left\{0,\,\frac{K}{2}\right\} \\ \frac{1}{K}\left(\frac{\pi^2}{6}+\cos\left(\frac{2\pi}{K}q\right)\right) & ,\mathrm{else} \end{cases} . \qquad (12)$$

The additional cosine in (12) results from the spectral correlation introduced by the Hann window used for the short-time Fourier analysis. A derivation of the variance of cepstral coefficients for arbitrary μ can be found in [10].

In general, the applied smoothing process should be unbiased, *i.e.* the variance of γ_k should be reduced without affecting its mean. As presented in [10], with the shape parameters μ and $\bar{\mu}$ known, this bias can be removed easily by applying

$$\overline{\gamma}_k(l) = \widetilde{\gamma}_k(l)\,\frac{\mu}{\bar{\mu}(l)}\,e^{\psi(\bar{\mu}(l))-\psi(\mu)}\,, \qquad (13)$$

where $\psi(\cdot)$ is the psi-function [13, (8.360)].

5. EVALUATION

In this section we compare the proposed SPP estimator to the approaches presented in [3, 4, 5]. For the estimator of Malah *et al.*, we implement the iterative approach used in the experimental results presented in [3]. For the estimator of Cohen and Berdugo, we set the recursive smoothing constant $\beta = 0.48$ in [4, (23)], as we only use a 50% segment overlap for the short-time Fourier analysis.

For the smoothing in (7) we choose α_q, as

$$\alpha_q(l) = \begin{cases} 0.2 & ,q\in\{0,...,2\}\backslash\mathbb{Q}_{\mathrm{pitch}} \\ 0.4 & ,q\in\{3,...,23\}\backslash\mathbb{Q}_{\mathrm{pitch}} \\ 0.997 & ,q\in\{24,...,256\}\backslash\mathbb{Q}_{\mathrm{pitch}} \\ 0.5 & ,q\in\mathbb{Q}_{\mathrm{pitch}} \end{cases} , \qquad (14)$$

where $\mathbb{Q}_{\mathrm{pitch}}$ are the cepstral coefficients that represent the fundamental period. $\mathbb{Q}_{\mathrm{pitch}}$ is found by searching for the maximum in the upper cepstrum, as detailed in [9]. If the found cepstral peak is lower than a threshold of 0.2 the respective signal segment is assumed to be unvoiced, and $\mathbb{Q}_{\mathrm{pitch}}$ is an empty set. Additionally, as in [9, (7)], we recursively smooth α_q over time with the smoothing constant $\beta = 0.9$. With the given α_q and assuming that the true, unknown SNR ranges from -10 dB to 25 dB, the optimal fixed *a priori* SNR is found to be $\xi_k = 9\,\mathrm{dB}$ [5]. This fixed *a priori* SNR $\xi_k = 9\,\mathrm{dB}$, the *a priori* SPP $\rho = 0.5$, the shape parameter $\bar{\mu}$ determined via (10), $\mu = 1$, and the smoothed and bias corrected *a posteriori* SNR (13) are then used in (5) and (1) to estimate the *a posteriori* SPP $\mathcal{P}_k$.

In Figure 1 (see last page) the resulting SPP estimates are shown. It can be seen that the estimator proposed in [3] does not yield SPP estimates close to zero in speech absence. This undesired behavior is overcome by the estimators [4, 5] and the proposed approach. The drawback of [4, 5] is that the spectral harmonics of the male speaker are not resolved. The proposed approach not only yields low SPP estimates in speech absence, but also resolves the spectral harmonics.

We evaluate the SPP estimators in terms of the measures for speech distortion (SD) and noise leakage (NL) introduced in [5]. The measure for speech distortion indicates the percentage of the speech energy that the SPP estimator misses and is thus related to the missed-hit rate. The measure for noise leakage indicates the percentage of the noise energy that is not attenuated by the SPP estimator and is thus related to the false-alarm rate. Furthermore, we quantify the segmental SNR improvement when the SPP estimate $\mathcal{P}_k$ is applied multiplicatively to noisy speech coefficients Y_k. We process 320 speech samples from dialect region 6 of [14] which are phonetically balanced and are from both male and female speakers. The speech is disturbed by white Gaussian noise, babble noise inside a crowded restaurant, and nonstationary traffic noise at a busy street, respectively. The experimental results for input segmental SNRs between -10 and 15 dB are given in Figure 2 (this page). It can be seen that the proposed estimator exhibits considerably less noise leakage than the estimators [3, 4, 5] while yielding similar or lower speech distortion than [4, 5]. The estimator in [3] exhibits an even lower speech distortion as it does not yield values close to zero in speech absence. Consequently, it gives a large noise leakage and results in a poor SNR improvement. Especially in babble noise, the proposed estimator clearly outperforms the competing estimators in terms of the segmental SNR improvement.

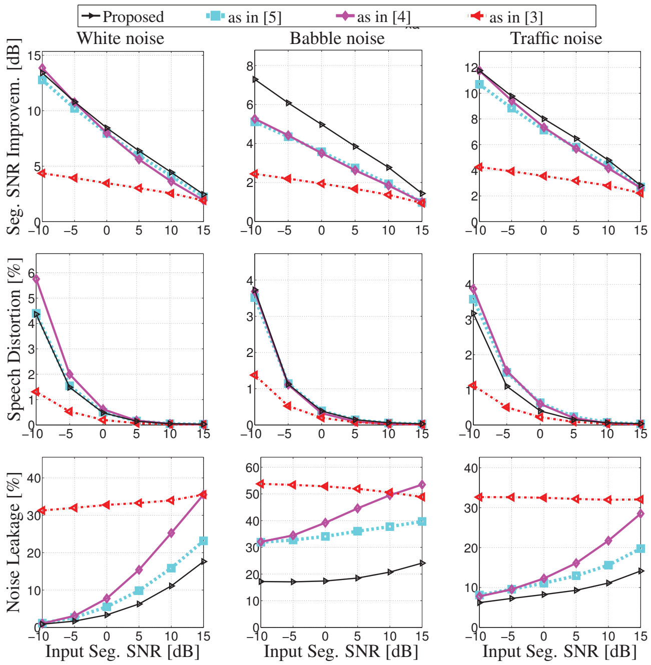
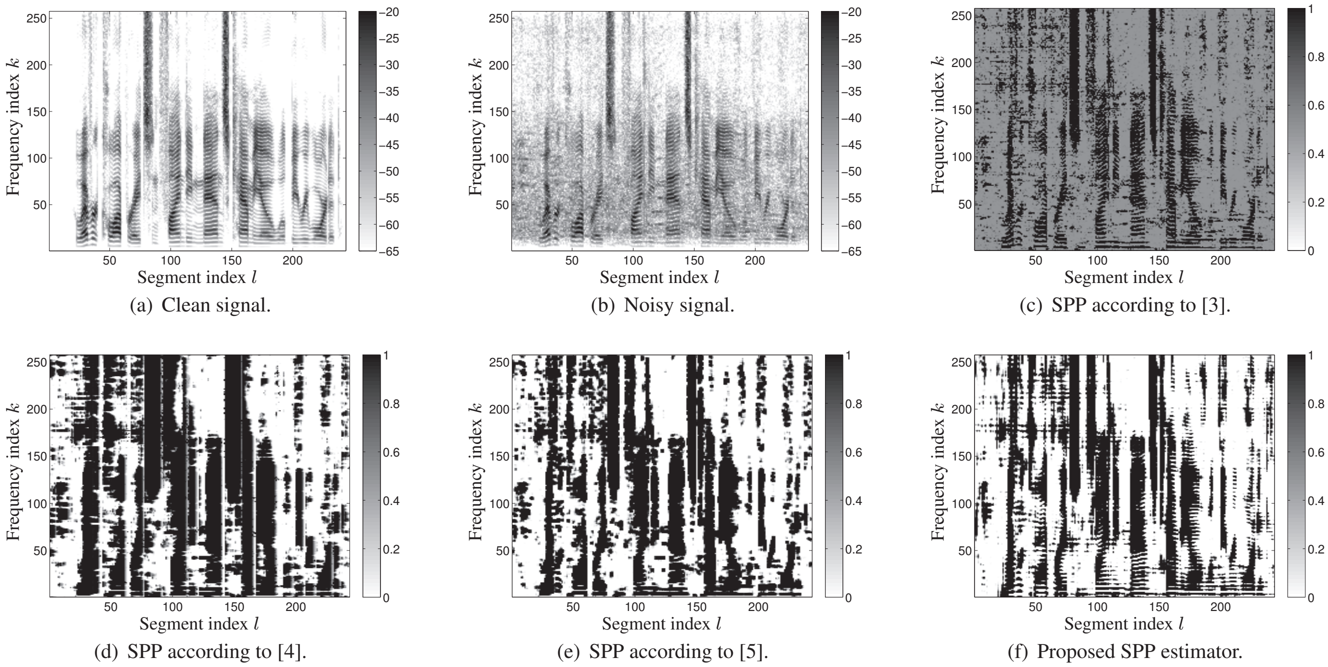


Fig. 2. The average segmental SNR improvement (top), speech distortion (middle), and noise leakage (bottom) averaged over 320 TIMIT sentences for white Gaussian noise (left), babble noise (middle), and nonstationary traffic noise (right).

6. CONCLUSION

In this paper, we have proposed an estimator for the speech presence probability (SPP) at each time-frequency point in the short-time Fourier domain, based on the temporal smoothing of the cepstrum. All required parameters are derived in an optimal way for a given set of cepstral smoothing parameters. As opposed to competing estimators, we use a fixed *a priori* SPP and a fixed *a priori* signal-to-noise ratio (SNR), *i.e.* they represent true *a priori* knowledge. While competing estimators combine SPP estimates based on local and global spectral smoothing, this is not necessary in the proposed estimator. The proposed approach achieves a higher frequency resolution, considerably less noise leakage, and a higher or similar SNR improvement, while obtaining lower or similar speech distortion as compared to the state-of-the-art estimators [4, 5].



(a) Clean signal.

(b) Noisy signal.

(c) SPP according to [3].

(d) SPP according to [4].

(e) SPP according to [5].

(f) Proposed SPP estimator.

Fig. 1. Clean speech (a), noisy speech (b), and the resulting SPP estimates (c)-(f) for speech from a male speaker disturbed by additive babble noise at 0 dB input segmental SNR. The signals in the spectrograms (a) and (b) have been pre-emphasized for a better visualization of high-frequency components. The estimator [3] in (c) does not yield SPP estimates close to zero in speech absence. The estimators in (d)-(f) overcome this undesired behavior. However, the estimators in (d) and (e) are not capable of resolving the spectral harmonics of the male speaker. This is clearly improved with the proposed approach in panel (f), resulting in less noise leakage without an increase in speech distortion.

7. REFERENCES

[1] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[3] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 789–792, 1999.

[4] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *ELSEVIER Signal Processing*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

[5] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

[6] T. Yu and J. H. L. Hansen, "A speech presence microphone array beamformer using model based speech presence probability estimation," *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 213–216, 2009.

[7] T. Gerkmann and R. Martin, "Soft decision combining for dual channel noise reduction," *ISCA Interspeech*, pp. 2134–2137, Sep. 2006.

[8] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[9] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4897–4900, Apr. 2008.

[10] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[11] I. Andrianakis and P. R. White, "MMSE speech spectral amplitude estimators with chi and gamma speech priors," *Proc. IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. III, pp. 1068–1071, 2006.

[12] R. Martin and T. Lotter, "Optimal recursive smoothing of non-stationary periodograms," *Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 167–170, Sep. 2001.

[13] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals Series and Products*, 6th ed., A. Jeffrey and D. Zwillinger, Ed. Academic Press, 2000.

[14] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST)*, 1988.