SINGLE-CHANNEL SPEECH PRESENCE PROBABILITY ESTIMATION USING INTER-FRAME AND INTER-BAND CORRELATIONS

Hajar Momeni[1,2,*], *Emanuël A. P. Habets*[1] *and Hamid Reza Abutalebi*[2]

[1]International Audio Laboratories Erlangen, Germany[†]
[2] Electrical and Computer Engineering Dept., Yazd University, Iran
h_momeni@stu.yazd.ac.ir, emanuel.habets@audiolabs-erlangen.de and habutalebi@yazd.ac.ir

ABSTRACT

The speech presence probability (SPP) plays an important role in many noise reduction and noise estimation methods. The SPP is commonly computed per time and frequency in the short time Fourier transform (STFT) domain based on the *a priori* speech absence probability and the *a priori* and *a posteriori* signal-to-noise ratios. Due to the STFT as well as the nature of the speech signal, there exists a correlation between subsequent time frames and neighboring frequency bands. In this work, we explicitly take these inter-frame and inter-band correlations into account when computing the SPP. The presented results demonstrate that we can increase the detection accuracy of the SPP estimator by taking a few neighboring time and frequency bins into account.

Index Terms— speech presence probability, inter-frame correlations, inter-band correlations.

1. INTRODUCTION

For many noise reduction and noise estimation methods, an estimator for the speech presence probability (SPP) in each time-frequency (TF) unit is of great interest. Clean-speech estimators, for example, are often derived under the assumption that speech is actually present. As this assumption is true neither during speech pauses nor between spectral bins of the harmonics of a voiced sound, the SPP should be taken into account [1–4]. Available noise power spectral density (PSD) estimators also make use of the SPP to decide when to update the noise PSD [5–7].

The SPP is commonly computed per TF unit in the short time Fourier transform (STFT) domain based on the *a priori* speech absence probability and the *a priori* and *a posteriori* signal-to-noise ratios. Most *a posteriori* speech presence probability (SPP) estimators are derived under the assumption that the spectral coefficients of the speech and noise can be modeled using complex Gaussian random variables. Moreover, it is commonly assumed that the time and frequency units are mutually uncorrelated across time and frequency. The spectral coefficients obtained after computing the STFT are both correlated across time and frequency. In addition, subsequent time frames are correlated due to the short-term stationarity of the speech signal, and neighboring frequency bins are correlated due to the harmonic structure of voiced speech segments [8].

In recent works [8–10], the inter-band correlations were explicitly used to derive novel noise reduction filters. In other works (c.f. [9, 11, 12]), the inter-frame correlations have been used to derive novel single and multichannel noise reduction filters. In [12], a single-channel noise reduction filter that uses the inter-frame correlations was derived that is able to reduce noise without distorting the desired speech. In [13], a fullband voice activity detector was proposed that takes the inter-band correlations into account. In [14], Gerkmann et al. noted that SPP estimators that rely on an observation of the noisy periodogram suffer from random fluctuations. Among other modifications, they proposed to compute an average of the *a posteriori* signal-to-noise ratio (SNR) under the assumption that the speech energy is distributed homogeneously over a small spectrogram region. Although the correlation that results from the spectral analysis is partly taken into account, the correlation due to the speech or noise signal is not taken into account.

In this paper, our goal is to estimate the narrowband SPP using a single noisy speech signal. In particular, we explicitly exploit the inter-frame and inter-band correlations when estimating the SPP in each TF unit. The obtained SPP estimator is similar to the one presented in [15] that was developed to exploit inter-channel correlations. In [15] a simplified SPP estimator was obtained under the implicit assumption that the correlation matrix of the desired signal is of rank one. Here, we investigate the performance of the SPP estimator with full rank and rank one assumptions. The presented results demonstrate that we can increase the detection accuracy of the SPP estimator by taking a few neighboring time and frequency bins into account.

The paper is organized as follows: in Section 2, the problem is formulated. In Section 3, the SPP estimator is derived that is able to take both inter-frame and inter-band correlations into account. In Section 4, the experimental results are provided and discussed. Finally, Section 5 concludes the paper.

2. PROBLEM FORMULATION

We consider the well-accepted signal model in which a microphone captures a desired signal that is corrupted by additive noise. In the short-time Fourier transform (STFT) domain we can express the spectral coefficients of the received signal at time-frame m and discrete-frequency k as

$$Y(k,m) = X(k,m) + V(k,m), \tag{1}$$

where $X(k,m)$ is the desired signal and $V(k,m)$ is the additive noise. We assume that the spectral coefficients $X(k,m)$ and $V(k,m)$ are uncorrelated and zero-mean complex Gaussian random variables.

Because of the properties of the STFT and the nature of the speech signal, it is likely that the TF unit of interest is correlated with

[†]A joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS, Germany.

[*]Ms Momeni was a Visiting Researcher at the AudioLabs from September 2013 till February 2014.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

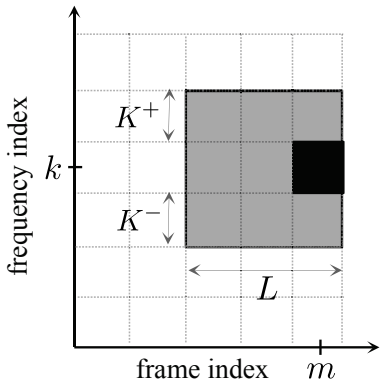


Fig. 1. Illustration of the time and frequency units that are taken into account to construct the input signal vector $\mathbf{y}(k,m)$. The black square indicates the current unit (k,m). The gray area around it indicates the additional TF units used.

neighboring TF units. For computing the SPP at TF unit (k,m), we take only a selected number of spectral coefficients into account as shown in Fig. 1.

We define an input signal vector, $\mathbf{y}(k,m)$, that contains all spectral coefficients that are taken into account as

$$\begin{aligned}\mathbf{y}(k,m) = \big[&Y(k-K^-,m)\cdots Y(k,m)\cdots Y(k+K^+,m)\\ &Y(k-K^-,m-1)\cdots Y(k,m-1)\cdots Y(k+K^+,m-1)\\ &\cdots Y(k-K^-,m-L+1)\cdots Y(k,m-L+1)\\ &\cdots Y(k+K^+,m-L+1)\big]^T,\end{aligned} \quad (2)$$

where L is the number of consecutive time frames used for each frequency bin[1], and K^- and K^+ are, respectively, the numbers of consecutive frequency bands before and after the kth bin used for each TF unit. In the following we assume that $K = K^- = K^+$ The input signal vector $\mathbf{y}(k,m)$ has a length $M = L\,(2\,K+1)$.

The vector $\mathbf{y}(k,m)$ can be expressed as

$$\mathbf{y}(k,m) = \mathbf{x}(k,m) + \mathbf{v}(k,m) \quad (3)$$

where

$$\begin{aligned}\mathbf{x}(k,m) = [&X(k-K,m)\\ &\cdots X(k,m)\cdots X(k+K,m-L+1)]^T\end{aligned} \quad (4)$$

and

$$\begin{aligned}\mathbf{v}(k,m) = [&V(k-K,m)\\ &\cdots V(k,m)\cdots V(k+K,m-L+1)]^T.\end{aligned} \quad (5)$$

The SPP is defined as the *a posteriori* probability that speech is present given the noisy observation and the statistical properties of the speech and the noise. Now, our objective is to estimate the *a posteriori* SPP at time frame m and frequency bin k, given the noisy input signal vector $\mathbf{y}(k,m)$.

[1]We can use different numbers of consecutive time-frames for different frequencies but to simplify the presentation, we use the same number L.

3. SPEECH PRESENCE PROBABILITY ESTIMATION

In this section, we describe how the *a posteriori* SPP can be computed. We first define the following hypotheses that indicate speech presence and absence:

$$\begin{aligned}\mathcal{H}_0 &: \ \mathbf{y}(k,m) = \mathbf{v}(k,m), \ \text{indicating speech absence}\\ \mathcal{H}_1 &: \ \mathbf{y}(k,m) = \mathbf{x}(k,m) + \mathbf{v}(k,m), \ \text{indicating speech presence.}\end{aligned}$$

Assuming that the speech and noise components are complex Gaussian random vectors with uncorrelated identically distributed real and imaginary parts, the likelihoods $p[\mathbf{y}\,|\,\mathcal{H}_0]$ and $p[\mathbf{y}\,|\,\mathcal{H}_1]$ can be written in closed form according to [15] as[2]

$$p[\mathbf{y}\,|\,\mathcal{H}_0] = \frac{1}{\pi^M \det[\mathbf{\Phi_v}]}\, e^{-\mathbf{y}^{\mathrm{H}}\mathbf{\Phi_v}^{-1}\mathbf{y}}, \quad (7)$$

$$p[\mathbf{y}\,|\,\mathcal{H}_1] = \frac{1}{\pi^M \det[\mathbf{\Phi_x}+\mathbf{\Phi_v}]}\, e^{-\mathbf{y}^{\mathrm{H}}(\mathbf{\Phi_v}+\mathbf{\Phi_x})^{-1}\mathbf{y}}, \quad (8)$$

where $\det[\cdot]$ denotes the determinant of a matrix and $\mathbf{\Phi_v} = E\{\mathbf{v}\mathbf{v}^H\}$ and $\mathbf{\Phi_x} = E\{\mathbf{x}\mathbf{x}^H\}$ denote the correlation matrices of the noise and speech, respectively.

The generalized likelihood ratio (GLR) is defined as

$$\Lambda = \frac{q}{1-q}\,\frac{p[\mathbf{y}\,|\,\mathcal{H}_1]}{p[\mathbf{y}\,|\,\mathcal{H}_0]}, \quad (9)$$

where $q = p[\mathcal{H}_1]$ denotes the *a priori* SPP.

The SPP can be obtained from the Bayes rule as follows

$$p[\mathcal{H}_1\,|\,\mathbf{y}] = \frac{\Lambda}{1+\Lambda} \quad (10a)$$

$$= \left\{1 + \frac{1}{\Lambda}\right\}^{-1}. \quad (10b)$$

Without making any further assumptions, GLR in (9) can be rewritten as [15]

$$\Lambda = \frac{q}{1-q}\,\frac{\det[\mathbf{\Phi_v}]}{\det[\mathbf{\Phi_x}+\mathbf{\Phi_v}]}\, e^{\mathbf{y}^{\mathrm{H}}[\mathbf{\Phi_v}^{-1}-(\mathbf{\Phi_v}+\mathbf{\Phi_x})^{-1}]\mathbf{y}}. \quad (11)$$

Although the SPP estimator is similar to [15], it should be noticed that we use the inter-frame and inter-band correlations instead of inter-channel correlations.

In [15], it is implicitly assumed that the speech correlation matrix $\mathbf{\Phi_x}$ is of rank one. Using the matrix inversion lemma, the GLR can then be written as

$$\Lambda = \frac{q}{1-q}\,\frac{1}{1+\mathrm{tr}\{\mathbf{\Phi_v}^{-1}\mathbf{\Phi_x}\}}\,\exp\left\{\frac{\mathbf{y}^{\mathrm{H}}\mathbf{\Phi_v}^{-1}\mathbf{\Phi_x}\mathbf{\Phi_v}^{-1}\mathbf{y}}{1+\mathrm{tr}\{\mathbf{\Phi_v}^{-1}\mathbf{\Phi_x}\}}\right\}, \quad (12)$$

where $\mathrm{tr}\{\cdot\}$ denotes the trace of a matrix. Introducing the quantities

$$\xi = \mathrm{tr}\{\mathbf{\Phi_v}^{-1}\mathbf{\Phi_x}\}, \quad (13)$$

$$\beta = \mathbf{y}^{\mathrm{H}}\mathbf{\Phi_v}^{-1}\mathbf{\Phi_x}\mathbf{\Phi_v}^{-1}\mathbf{y}, \quad (14)$$

the SPP can be expressed as

$$p[\mathcal{H}_1\,|\,\mathbf{y}] = \left\{1 + \frac{1-q}{q}\,[1+\xi]\exp\left[-\frac{\beta}{1+\xi}\right]\right\}^{-1}. \quad (15)$$

For $L = 1$ and $K = 0$, the SPP estimator reduces to the traditional single-channel SPP estimator [2].

[2]The time and frequency indices are omitted for brevity.

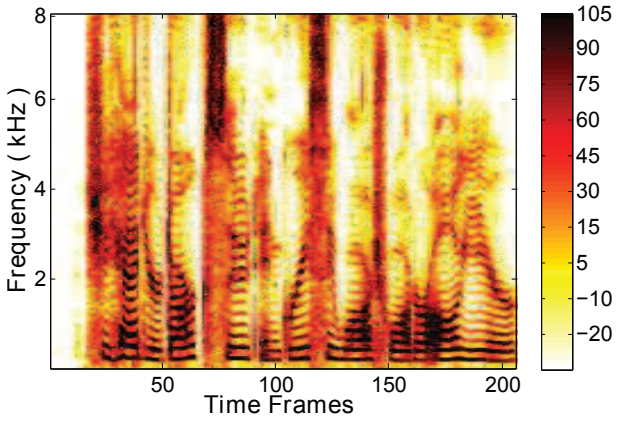


Fig. 2. Spectrogram of the clean female speech sample: "She had your dark suit in greasy wash water all year".

4. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the narrowband SPP estimator for four different scenarios: In *Scenario 1*, we do not use any other TF units (i.e., $L = 1$ and $K = 0$). In *Scenario 2*, we use only inter-frame correlations (i.e., $L > 1$ and $K = 0$). In *Scenario 3*, we only use inter-band correlations (i.e., $L = 1$ and $K > 0$). In *Scenario 4*, we use both the inter-band and inter-frame correlations (i.e., $L > 1$ and $K > 0$). Scenarios 2-4 have, to the best of our knowledge, not been investigated before.

4.1. Experimental Setup

In the experiments, clean speech samples were used from the TIMIT database [16]. The sampling frequency was 16 kHz. The additive noise consisted of white Gaussian noise and different input SNRs were obtained by changing the level of the noise. The STFT is computed using a 32 ms hamming window with 50% overlap. The correlation matrix of the noisy signal, $\mathbf{y}(k,m)$, is computed recursively using

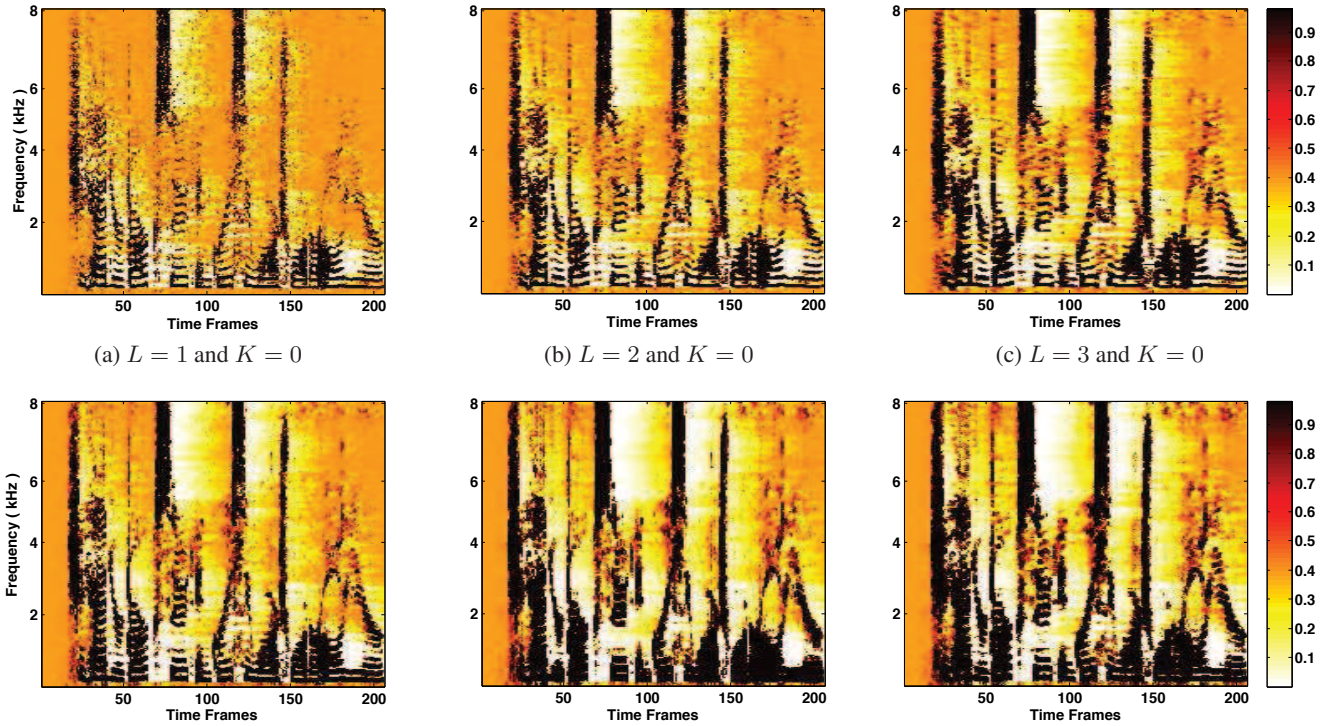
$$\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m) = \alpha_y \, \widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m-1) + (1-\alpha_y)\, \mathbf{y}(k,m)\mathbf{y}^H(k,m), \quad (16)$$

where α_y denotes the forgetting factor that was set to 0.85. Note that even though the noise is white, the correlation matrix is not an identity matrix due to overlapping analysis windows of the STFT.

Since our objective is to analyze the performance of the SPP estimator, the noise correlation matrix is computed from the first second during which the speech was absent such that $\widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,m) = \widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m)$. The correlation matrix of the clean speech was computed using $\widehat{\mathbf{\Phi}}_{\mathbf{x}}(k,m) = \mathcal{P}\{\widehat{\mathbf{\Phi}}_{\mathbf{y}}(k,m) - \widehat{\mathbf{\Phi}}_{\mathbf{v}}(k,m)\}$, where $\mathcal{P}\{\cdot\}$ is an operation that sets all negative eigenvalues to zero to ensure that the resulting matrix is positive definite. The *a priori* SPP was set to $q(k,m) = 0.4$ as used in [15].

4.2. Examination of Four Scenarios

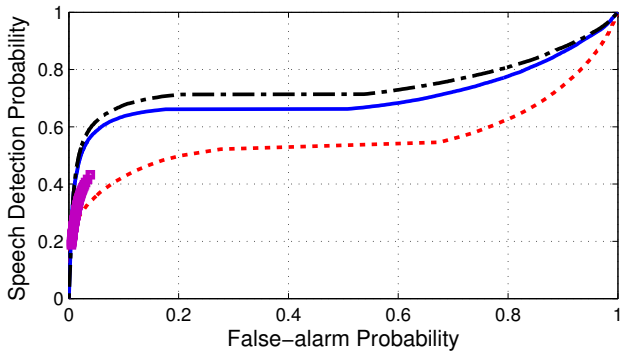
Here we examine the estimated SPP obtained for the four different scenarios using a female speech sample shown in Fig. 2. In Fig. 3, the estimated SPP is shown for different values of L and K. The result shown in Fig. 3(a), is obtained using the traditional SPP esti-



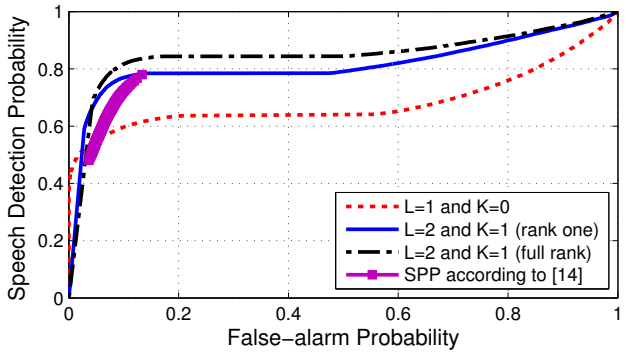
(a) $L = 1$ and $K = 0$ (b) $L = 2$ and $K = 0$ (c) $L = 3$ and $K = 0$

(d) $L = 1 = K = 1$ (e) $L = 1$ and $K = 2$ (f) $L = 2$ and $K = 1$

Fig. 3. Time and frequency dependent SPP, input SNR=15 dB: (a) Scenario 1: $L = 1$ and $K = 0$, (b) Scenario 2: $L = 2$ and $K = 0$, (c) Scenario 2: $L = 3$ and $K = 0$, (d) Scenario 3: $L = 1$ and $K = 1$, (e) Scenario 3: $L = 1$ and $K = 2$, (f) Scenario 4: $L = 2$ and $K = 1$.



(a) Input SNR = 5 dB



(b) Input SNR = 15 dB

Fig. 4. ROC curves for two different input SNRs.

mator that does not take the inter-frame and inter-band correlations into account. Comparing the estimated SPP in Fig. 3(a) to Fig. 2, we can observe that the SPP is not always detected correctly. By using multiple time frames, we can achieve better detection at frequencies above 3 kHz (see Fig. 3(b) for $L = 2$ and Fig. 3(c) for $L = 3$ and $K = 0$). Even better results are obtained by using $L = 1$ and multiple frequency bins as shown in Fig. 3(d-e). For $L = 1$ and $K = 2$, we notice that the fine structure of the speech is no longer preserved while the unvoiced speech segments are better preserved. A good compromise is obtained using $L = 2$ and $K = 1$ as shown in Fig. 3(f).

4.3. Numerical Results

The performance of the proposed SPP estimator was evaluated by examining the speech detection and false-alarm probabilities (P_{d} and P_{fa}) for Scenarios 1 and 4. For this experiment, one male and one female speech sample were concatenated. We define P_{d} as the ratio of correct speech decisions to the speech TF units, and P_{fa} as that of false speech decisions to the non-speech TF units. The speech or non-speech TF units were obtained by comparing whether the power of a TF unit of the clean speech was larger or smaller than a predefined level that was set to -60 dB below the maximum instantaneous power across all TF units. The speech detection, P_{d}, was computed as the ratio of the total number of speech TF units with a SPP larger or equal than a given threshold (between zero and one), to the total number of speech TF units. The false-alarm probability, P_{fa}, was computed similar to P_{d} using the non-speech TF units instead of the speech TF units. For any given input SNR and threshold between zero and 1, we computed both P_{d} and P_{fa}. The receiver operating characteristic (ROC) curve that shows the tradeoff between P_{d} and P_{fa} is then obtained by plotting both values [13].

	Input SNR		
	5 dB	10 dB	15 dB
$L = 1$ and $K = 0$	0.329	0.397	0.512
$L = 2$ and $K = 1$ (rank one)	0.540	0.581	0.614
$L = 2$ and $K = 1$ (full rank)	**0.572**	**0.615**	**0.658**
SPP according to [14]	0.434	0.465	0.467

Table 1. Speech detection probability for different input SNRs for a false-alarm probability of 0.05.

Here we tested four SPP estimators: i) computed using (15) (i.e., with $L = 1$ and $K = 0$), ii) computed using (10) and (12) with $L = 2$ and $K = 1$ (i.e., with the rank one assumption), iii) computed using (10) and (11) with $L = 2$ and $K = 1$ (i.e., with full rank assumption), and iv) the SPP estimator proposed in [14] using the reported parameters. The ROC curves of the four SPP estimators for an input SNR of 5 and 15 dB are shown in Figs 4(a) and 4(b), respectively. Thresholds were set from a value close to one, to zero with decreasing steps of 0.001. In Figs. 4(a) and 4(b), it can be seen that we can achieve a higher speech detection probability for a given false-alarm probability using both inter-frame and inter-band correlations. Furthermore, the SPP estimator that uses both inter-frame and inter-band correlations with full rank assumption outperforms the estimator with the rank one assumption as well as the estimator proposed in [14].

Finally, we show the speech detection probability, P_{d}, as a function of the input SNRs for a given false-alarm probability, P_{fa}, was set to 0.05. The results for an input SNR of 5, 10, and 15 dB are depicted in Table. 1. Exploiting the inter-frame and inter-band correlations (as done in Scenario 4) results in a higher speech detection probability compared to the traditional SPP estimator (as done in Scenario 1) and the estimator proposed in [14] for all SNRs. We can notice that the difference between the detection probabilities of the estimator with rank one and with full rank assumptions (Scenario 4) monotonically increases with the input SNRs.

5. CONCLUSIONS

We proposed a general single-channel SPP estimator that can explicitly exploit inter-frame and inter-band correlations. As a special case, we obtained the traditional SPP estimator that is computed based on only the current noisy observation and an estimate of the clean speech and noise statistics. For the evaluated signals, it was shown that the detection accuracy of the SPP estimator can be increased by using inter-frame and/or inter-band correlations and as such consistently outperforms the traditional SPP estimator and the estimator proposed in [14]. Especially at low input SNRs, which is often the case at high frequencies, the additional information improves the detection accuracy. It is expected that the use of the proposed narrowband SPP estimator can further improve the performance of single-channel noise reduction filters as well single-channel noise PSD estimators as their performance depends to a large extent on the accuracy of the SPP estimator.

6. REFERENCES

[1] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 2, pp. 137–145, Apr. 1980.

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

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

[4] A. Abramson and I. Cohen, "Simultaneous detection and estimation approach for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2348–2359, Nov. 2007.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[6] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2011, pp. 145–148.

[7] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, 2011.

[8] E. Plourde and B. Champagne, "Multidimensional STSA estimators for speech enhancement with correlated spectral components," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3013–3024, Jul. 2011.

[9] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, ser. SpringerBriefs in Electrical and Computer Engineering. Springer-Verlag, 2011.

[10] J. Chen and J. Benesty, "Single-channel noise reduction in the STFT domain based on the bifrequency spectrum," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 97–100.

[11] E. A. P. Habets, "A distortionless subband beamformer for noise reduction in reverberant environments," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel Aviv, Israel, Aug. 2010, pp. 1–4.

[12] Y. A. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1256–1269, 2012.

[13] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

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

[15] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[16] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Technical Report, Dec. 1988.