CONDITIONAL MMSE-BASED SINGLE-CHANNEL SPEECH ENHANCEMENT USING INTER-FRAME AND INTER-BAND CORRELATIONS

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

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

ABSTRACT

Obtaining an estimate of clean speech for each time-frequency (TF) unit continues to be of importance in single-channel speech enhancement. Recently, it has been proposed to exploit inter-frame and inter-band correlations in a variety of speech processing applications. To estimate the clean speech, we propose in this contribution a conditional minimum mean squared error (MMSE)-based filter which exploits both inter-frame and inter-band correlations and takes into account the speech presence uncertainty. The speech presence uncertainty is provided by a recently proposed *a posteriori* speech presence probability (SPP) estimator that can also take into account the inter-frame and inter-band correlations. Simulation results demonstrate that the conditional MMSE-based filter in combination with the previously proposed SPP estimator and a fixed *a priori* SPP results in less distorted speech compared to the other SPP estimators.

Index Terms— Inter-frame and inter-band correlations, single channel noise reduction, speech presence probability.

1. INTRODUCTION

Speech enhancement is commonly performed in the short-time Fourier transform (STFT) domain. In the last three decades, the clean speech in each time-frequency (TF) unit is commonly estimated from noisy speech in the same TF unit (c.f., [1]). A common assumption in the derivation of clean speech estimators is that speech TF units are mutually uncorrelated across time and frequency. Considering the STFT, the short-term stationarity of the speech signal, and the harmonic structure of voiced speech segments [2], researchers have recently started to exploit the correlation between adjacent TF units. In [3–5], optimal noise reduction filters were for example proposed that exploit the inter-frame correlations. Moreover, the inter-band correlations have been explicitly used to derive novel noise reduction filters in [2,4,6]. The inter-frame and inter-band correlations has been recently studied in the context of single-channel *a posteriori* speech presence probability (SPP) estimation [7], blind speech separation application [8,9], stereophonic acoustic echo suppression [10], and multichannel noise reduction [11].

Utilizing the speech presence uncertainty in the estimation of the clean speech [1, 12–16] or noise statistics [17, 18] has a long history. The SPP-based clean-speech estimators are structured based on incorporating the SPP into the estimator [1, 19]. In previous works it was shown that if the SPP estimator is able to better detect the speech TF units, SPP-based clean speech estimators as well as noise estimators can achieve a higher performance. In [12], Gerkmann *et al.* proposed to compute an average of the *a posteriori* SNR under the assumption that the speech energy is distributed homogeneously over a TF region. They then derived an SPP estimator from the averaged *a posteriori* SNR. In [7], the present authors proposed a single-channel SPP estimator that resulted in a higher probability of speech detection for a given false alarm rate using both inter-frame and inter-band correlations.

In this paper, we derive an MMSE-based filter for single-channel noise reduction that takes the inter-frame and inter-band correlations into account. To include the speech presence uncertainty, we formulate a conditional MMSE-based estimator based on the *a posteriori* SPPs. The *a posteriori* SPP is estimated using the inter-frame and inter-band correlations of the TF units surrounding the TF unit of interest. Moreover, we examine two methods to determine the required *a priori* SPP. Finally, we propose a procedure similar to that in [12] by setting the fixed priors to obtain SPP estimates close to one/zero for speech/non-speech TF units.

The remaining sections are structured as follows: In Section 2, the chosen model for signals is introduced. In Section 3, a conditional MMSE-based filter that exploits both inter-frame and inter-band correlations is derived. In Section 4, the SPP estimator is reviewed. In Section 5, the performance of the proposed filter is evaluated. In Section 6, the paper is concluded.

2. SIGNAL MODEL

We consider the well-known signal model in which a microphone captures the desired signal that is corrupted by additive noise. In the 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 neighboring TF units. To take this information into account, we adopt the signal model proposed in [7].

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

$$\mathbf{y}(k,m) = \Big[\mathbf{c}^{\mathrm{T}}(k,m)\,\mathbf{c}^{\mathrm{T}}(k,m-1)\,\ldots\,\mathbf{c}^{\mathrm{T}}(k,m-L+1)\Big]^{\mathrm{T}}, \tag{2}$$

[†]A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

with

$$\mathbf{c}(k,m) = \left[Y(k-K^-,m) \cdots Y(k,m) \cdots Y(k+K^+,m)\right]^{\mathrm{T}}, \quad (3)$$

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. The input signal vector $\mathbf{y}(k,m)$ has a length $M = L\,(K^- + K^+ + 1)$.

When speech is present with certainty, the vector $\mathbf{y}(k,m)$ can be expressed as

$$\mathbf{y}(k,m) = \mathbf{x}(k,m) + \mathbf{v}(k,m). \quad (4)$$

Our objective is to estimate the desired signal $X(k,m)$ by applying a filter

$$\mathbf{h}(k,m) = [H_0(k,m)\, H_1(k,m) \cdots H_{M-1}(k,m)]^{\mathrm{T}} \quad (5)$$

to the input signal vector $\mathbf{y}(k,m)$. The filter exploits the inter-frame and inter-band correlations, and takes into account the speech presence uncertainty. An estimate of the desired signal $X(k,m)$ is then given by

$$\begin{aligned} Z(k,m) &= \sum_{i=0}^{M-1} H_i^*(k,m) Y_i(k,m) \\ &= \mathbf{h}^{\mathrm{H}}(k,m)\mathbf{y}(k,m), \end{aligned} \quad (6)$$

where $Y_i(k,m)$ is the ith element of the vector $\mathbf{y}(k,m)$. The superscripts * and $^{\mathrm{H}}$ are conjugate and hermitian operators, respectively. Let $\mathbf{\Phi_y}(k,m) = E\left[\mathbf{y}(k,m)\mathbf{y}^{\mathrm{H}}(k,m)\right]$, $\mathbf{\Phi_x}(k,m) = E\left[\mathbf{x}(k,m)\mathbf{x}^{\mathrm{H}}(k,m)\right]$ and $\mathbf{\Phi_v}(k,m) = E\left[\mathbf{v}(k,m)\mathbf{v}^{\mathrm{H}}(k,m)\right]$ respectively denote the correlation matrices of the noisy speech, clean speech, and noise, where $E[\cdot]$ denotes the mathematical expectation.

3. CONDITIONAL MMSE-BASED FILTER USING INTER-FRAME AND INTER-BAND CORRELATIONS

In traditional single-channel noise reduction systems, the enhanced TF unit, $Z(k,m)$, is commonly the product of a gain, $H_0(k,m)$, and the noisy observation at each TF unit, i.e., $Z(k,m) = H_0(k,m)\, Y(k,m)$. By considering the complex Gaussian distribution models for speech and noise and assuming mean squared error (MSE) between $Z(k,m)$ and $X(k,m)$, the form of Wiener filter can be obtained so that the gain, $H_0(k,m)$, is the ratio of the variance of the clean speech over the variance of the noisy speech, i.e., $H_0(k,m) = H_{\mathrm{W}}(k,m) = \phi_X(k,m)/(\phi_X(k,m) + \phi_V(k,m))$ where $\phi_X(k,m)$ and $\phi_V(k,m)$ are the variances of clean speech and noise at time frame m and frequency bin k, respectively.

To include the speech presence uncertainty, we first define the following speech absence and presence hypotheses:

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

Considering the speech absence and presence hypotheses, a conditional MMSE estimate of $X(k,m)$ can be obtained from a noisy observation $\mathbf{y}(k,m)$ as follows:

$$\begin{aligned} Z(k,m) &= E\left[X(k,m)|\mathbf{y}(k,m)\right] \\ &= p(k,m)\, E\left[X(k,m)|\mathbf{y}(k,m), \mathcal{H}_1(k,m)\right] \\ &\quad + (1-p(k,m))\, E\left[X(k,m)|\mathbf{y}(k,m), \mathcal{H}_0(k,m)\right], \end{aligned} \quad (8)$$

where $E\left[X(k,m)|\mathbf{y}(k,m), \mathcal{H}_1(k,m)\right]$ is the estimated desired signal when speech is present, and $p(k,m)$ is the *a posteriori* SPP that is described in Section 4. When speech is absent, a gain $H_{\min}(k)$ ($H_{\min}(k) \ll 1$) is applied to ensure that the residual noise in the output signal sounds natural. Hence, $E\left[X(k,m)|\mathbf{y}(k,m), \mathcal{H}_0(k,m)\right]$ is equal to $H_{\min}(k)\, Y(k,m)$.

It is known that considering the MMSE criteria, $E\left[X(k,m)|\mathbf{y}(k,m), \mathcal{H}_1(k,m)\right]$ can be obtained using the Wiener filter $\mathbf{h}_{\mathrm{W}}(k,m)$ [20]:

$$\begin{aligned} \mathbf{h}_{\mathrm{W}}(k,m) &= \mathbf{\Phi_y}^{-1}(k,m)\mathbf{\Phi_x}(k,m)\mathbf{i}_{K^-+1} \\ &= \left[\mathbf{\Phi_x}(k,m) + \mathbf{\Phi_v}(k,m)\right]^{-1}\mathbf{\Phi_x}(k,m)\mathbf{i}_{K^-+1} \\ &= \left[\mathbf{I} - \mathbf{\Phi_y}^{-1}(k,m)\mathbf{\Phi_v}(k,m)\right]\mathbf{i}_{K^-+1}, \end{aligned} \quad (9)$$

where $\mathbf{I}$ is the identity matrix of size M, and $\mathbf{i}_{K^-+1}$ is the $(K^- + 1)$th column of $\mathbf{I}$. In the following, we assume $K = K^- = K^+$. It should be noted that the computational complexity increases when L and/or K increase.

Taking into account the speech presence uncertainty, the output signal can then be computed using:

$$\begin{aligned} Z(k,m) = p(k,m)\, &\mathbf{h}_{\mathrm{W}}{}^{\mathrm{H}}(k,m)\mathbf{y}(k,m) \\ &+ (1-p(k,m))\, H_{\min}(k)\, Y(k,m). \end{aligned} \quad (10)$$

For $L = 1$ and $K = 0$, the $\mathbf{h}(k,m)$ reduces to the traditional Wiener filter $H_{\mathrm{W}}(k,m)$ such that

$$\begin{aligned} Z(k,m) = [p(k,m)\, &H_{\mathrm{W}}(k,m) \\ &+ (1-p(k,m))\, H_{\min}(k)]\; Y(k,m). \end{aligned} \quad (11)$$

4. SPEECH PRESENCE PROBABILITY ESTIMATION

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, the objective is to estimate the SPP at time frame m and frequency bin k given the noisy input signal vector $\mathbf{y}(k,m)$. In this section, we describe how the SPP can be computed by considering the inter-frame and inter-band correlations of TF unit of interest [7].

Assuming that the speech and noise components are complex Gaussian random vectors with uncorrelated identically distributed real and imaginary parts and considering the speech absence and presence hypotheses, the likelihoods $f[\mathbf{y}(k,m)\,|\,\mathcal{H}_0(k,m)]$ and $f[\mathbf{y}(k,m)\,|\,\mathcal{H}_1(k,m)]$ can be written in closed form as [21]

$$\begin{aligned} f[\mathbf{y}(k,m)\,|\,\mathcal{H}_0(k,m)] = &\frac{1}{\pi^M \det[\mathbf{\Phi_v}(k,m)]} \\ &\times \mathrm{e}^{-\mathbf{y}^{\mathrm{H}}(k,m)\mathbf{\Phi_v}^{-1}(k,m)\mathbf{y}(k,m)}, \end{aligned} \quad (12)$$

and

$$\begin{aligned} f[\mathbf{y}(k,m)\,|\,\mathcal{H}_1(k,m)] = &\frac{1}{\pi^M \det[\mathbf{\Phi_x}(k,m) + \mathbf{\Phi_v}(k,m)]} \\ &\times \mathrm{e}^{-\mathbf{y}^{\mathrm{H}}(k,m)(\mathbf{\Phi_v}(k,m)+\mathbf{\Phi_x}(k,m))^{-1}\mathbf{y}(k,m)}, \end{aligned} \quad (13)$$

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

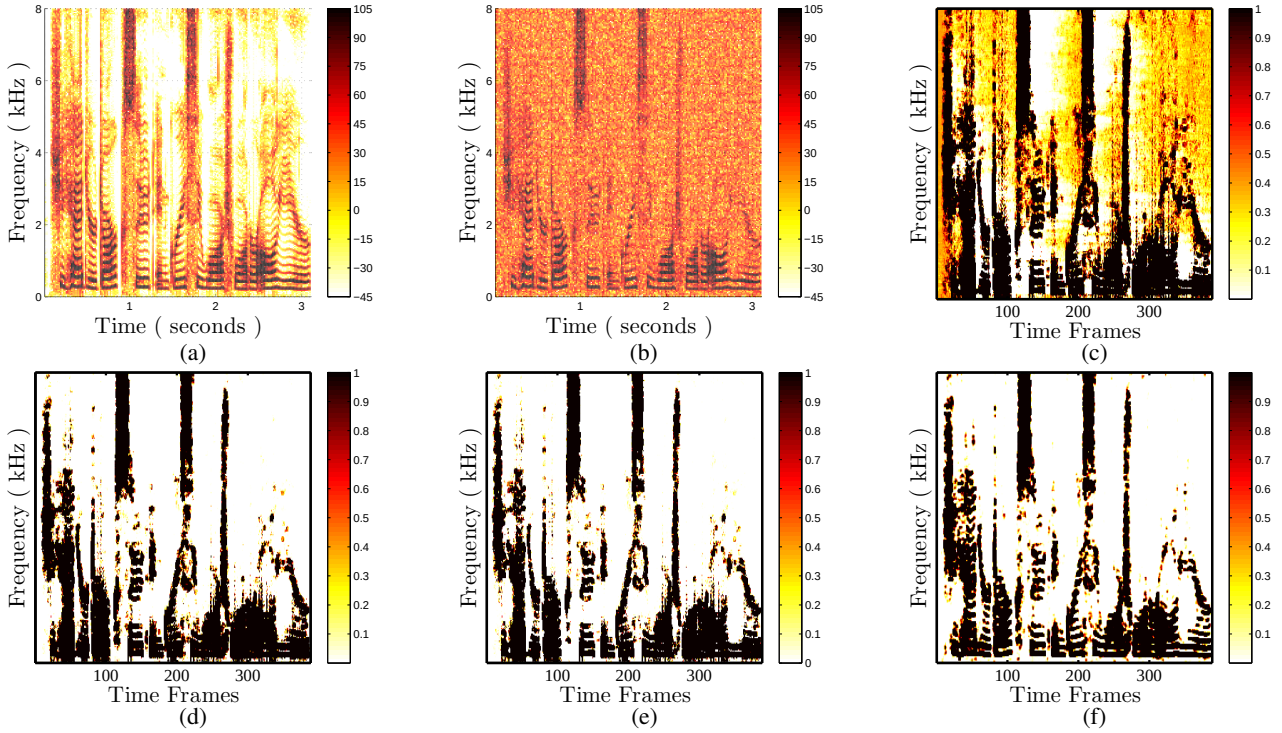

Fig. 1. Time and frequency dependent SPP, input SNR=15 dB: (a) Spectrogram of the clean speech, (b) Spectrogram of the noisy speech, (c) SPPs of "L2-qfix", (d) SPPs of "L2-zetafix"(e) SPPs of "L2-qCohen", (f) SPPs of "L1-Gerkmann".

where det[·] denotes the determinant of a matrix. The generalized likelihood ratio (GLR) is defined as [1]

$$\Lambda(k,m) = \frac{q(k,m)}{1-q(k,m)} \, \frac{f[\mathbf{y}(k,m) \,|\, \mathcal{H}_1(k,m)]}{f[\mathbf{y}(k,m) \,|\, \mathcal{H}_0(k,m)]}, \qquad (14)$$

where $q(k,m) = f[\mathcal{H}_1(k,m)]$ denotes the *a priori* SPP. Using (12) and (13), the GLR can be written as

$$\Lambda(k,m) = \frac{q(k,m)}{1-q(k,m)} \, \frac{\det[\mathbf{\Phi_v}(k,m)]}{\det[\mathbf{\Phi_x}(k,m)+\mathbf{\Phi_v}(k,m)]} \times \mathrm{e}^{\mathbf{y}^{\mathrm{H}}(k,m)[\mathbf{\Phi_v}^{-1}(k,m)-(\mathbf{\Phi_v}(k,m)+\mathbf{\Phi_x}(k,m))^{-1}]\mathbf{y}(k,m)}. \qquad (15)$$

Finally, the SPP is obtained from Bayes rule as follows [7]

$$p(k,m) \triangleq f[\mathcal{H}_1(k,m) \,|\, \mathbf{y}(k,m)] = \frac{\Lambda(k,m)}{1+\Lambda(k,m)}. \qquad (16)$$

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

$$p(k,m) = \left\{ 1 + \frac{1-q(k,m)}{q(k,m)} \, [1+\xi(k,m)] \, \mathrm{e}^{\left[-\frac{\gamma(k,m)\xi(k,m)}{1+\xi(k,m)}\right]} \right\}^{-1}, \qquad (17)$$

where $\gamma(k,m) = \frac{|Y(k,m)|^2}{\phi_V(k,m)}$ and $\xi(k,m) = \frac{\phi_X(k,m)}{\phi_V(k,m)}$ are *a posteriori* and *a priori* SNRs, respectively.

Finally, we have to determine the *a priori* SPP, $q(k,m)$, in (15). In this paper, we examine two different approaches to determine the *a priori* SPP:

1) Choosing a fixed value for $q(k,m)$, i.e., $q(k,m) = 0.4$ for all m and k as proposed in [7, 21].

2) Determining $q(k,m)$ as proposed in [19], i.e., by comparing the *a posteriori* SNR at each TF unit with predetermined thresholds. This estimator then leads to an SPP close to one/zero for speech/non-speech TF units.

In [12], the authors proposed an alternative approach to compute the single-channel *a posteriori* SPP. First they showed that the speech presence and absence likelihoods become identical if the *a priori* SNR is very small, which results in an SPP of 0.5. To mitigate this problem, they propose to compute the SPP using a fixed *a priori* SNR rather than a signal-dependent *a priori* SNR. The constant *a priori* SNR results in nonidentical speech presence and absence likelihoods when the *a priori* SNR is low. In [12], the fixed priors $q(k,m) = 0.5$ and $10\log_{10}\xi(k,m) = 8$ dB are used. It should be noted that also in our case the likelihoods $f[\mathbf{y}(k,m) \,|\, \mathcal{H}_0(k,m)]$ and $f[\mathbf{y}(k,m) \,|\, \mathcal{H}_1(k,m)]$ in (12) and (13) would be identical if the $\mathbf{\Phi_x}(k,m)$ contains small values. Therefore, we propose a procedure similar to that in [12] to obtain SPP estimates close to one/zero for speech/non-speech TF units. To overcome the identity of the two likelihoods in non-speech TF units, we set $\mathbf{\Phi_x}(k,m)$ to a factor of the $\mathbf{\Phi_v}(k,m)$, i.e., $\mathbf{\Phi_x}(k,m) = \rho \, \mathbf{\Phi_v}(k,m)$ where ρ is a fixed value. Moreover, the fixed *a priori* SPP was set to $q(k,m) = 0.5$.

5. PERFORMANCE EVALUATION

In the following performance evaluation, clean speech samples from the TIMIT database [22] were used. The speech signals were sampled at a frequency of 16 kHz. Three different noise types, consisted of white Gaussian, Factory 1, and babble, were examined at different input SNRs. The STFT was computed using a 32 ms Hamming window with 75% overlap. The correlation matrix of the noisy and noise signals were computed recursively from these signals using a forgetting factor of 0.92. The effect of the errors

in the estimated noise correlation matrix is a topic of future research. 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. Moreover, $H_{\min}(k)$ was set to -9 dB.

In this study the following algorithms were evaluated and compared:

- "L2-qfix": $L = 2$, $K = 1$, $Z(k,m)$ using (10), $p(k,m)$ using (15) and (16), $q(k,m) = 0.4$.
- "L2-zetafix": $L = 2$, $K = 1$, $Z(k,m)$ using (10), $p(k,m)$ using (15) and (16), $q(k,m) = 0.5$, $\mathbf{\Phi}_{\mathbf{x}}(k,m) = \rho\, \mathbf{\Phi}_{\mathbf{v}}(k,m)$ with $\rho = 7.94$ such that the SNR $\mathrm{tr}\{\mathbf{\Phi}_{\mathbf{v}}^{-1}(k,m)\mathbf{\Phi}_{\mathbf{x}}(k,m)\}$ is equal to 9 dB, which is similar to the *a priori* SNR used in [12]. Here $\mathrm{tr}\{\cdot\}$ indicates the trace of a matrix.
- "L2-qCohen": $L = 2$, $K = 1$, $Z(k,m)$ using (10), $p(k,m)$ using (15) and (16). The *a priori* SPP $q(k,m)$ was computed using the algorithm in [19].
- "L1-Gerkmann" [12]: $L = 1$, $K = 0$, $Z(k,m)$ using (11), $p(k,m)$ using the algorithm in [12] with $q(k,m) = 0.5$ and $10\log_{10}\xi(k,m) = 8$ dB.
- "L1-qCohen" [19]: $L = 1$, $K = 0$, $Z(k,m)$ using (11), $p(k,m)$ using(17), and $q(k,m)$ using the algorithm in [19].

First, we examine the estimated SPPs obtained from different algorithms. For this purpose, we used a female speech sample from the TIMIT database [22] shown in Fig. 1(a). The clean speech degraded by white Gaussian noise with an input SNR = 15 dB is depicted in Fig. 1(b). For those TF units where the speech is absent, it can be observed that the estimated SPPs shown in Fig. 1(d)-(f) are closer to zero compared to the SPP for "L2-qfix" shown in Fig. 1(c). However, when speech is present, the SPP for "L2-qfix" shown in Fig. 1(c) yields a higher probability compared to the SPPs shown in Fig. 1(d)-(f).

Secondly, we evaluate the performance of different SPP estimators, in terms of speech distortion (SD) and noise leakage (NL), which can act as measures for missed detection and false-alarm rate, respectively. The definitions of SD and NL are presented in [12]. In Fig. 2 the results for three different noise types and seven different input SNRs are depicted. The results are obtained by averaging on 16 sentences from the TIMIT database [22] (8 male, 8 female). It is observed that the "L2-qfix" yields the lowest SD and the highest NL values. The estimated SPPs close to $q(k,m)$ (rather than zero) in silent time frames resulted in high NL values. The "L1-qCohen" achieves the highest SD and the lowest NL values. The performance of the remaining algorithms lies in the middle.

In Fig. 2, we have also compared the performance of the above mentioned algorithms in speech enhancement application in various noise conditions in terms of the improvements in seg. SNR and PESQ. The seg. SNR measures the noise reduction and PESQ measures the overall speech quality [23]. In terms of the noise reduction, the evaluated algorithms can be ordered as follows:

"L1-qCohen"$\lessdot$"L1-Gerkmann"$\lessdot$"L2-qCohen"$\leq$"L2-zetafix"$\lessdot$"L2-qfix"

We can conclude that "L1-Gerkmann" outperforms "L1-qCohen", and that the proposed algorithms that exploit both inter-frame and inter-band correlations achieve a higher performance compared to the traditional ones. The superior performance of the proposed algorithm ("L2-qfix") may be explained by the lower SD values that are more important compared to higher NL values. Informal listening tests supported these findings.

6. CONCLUSIONS

We derived a conditional MMSE-based filter for single-channel noise reduction that takes the inter-frame and inter-band correlations into account and the speech presence uncertainty. The speech presence uncertainty was incorporated using the SPP, which could also be estimated using the inter-frame and inter-band correlations of the TF units surrounding the TF unit of interest. Different combinations of the conditional MMSE-based filter and SPP estimators were evaluated and compared in an experimental study. The experimental results showed that the conditional MMSE-based filter in combination with the SPP estimator that uses a fixed *a priori* SNR and exploits inter-frame and inter-band correlations resulted in the least amount of speech distortion, and highest segmental SNR and PESQ scores. A topic of future research is the estimation of the noise covariance matrix.

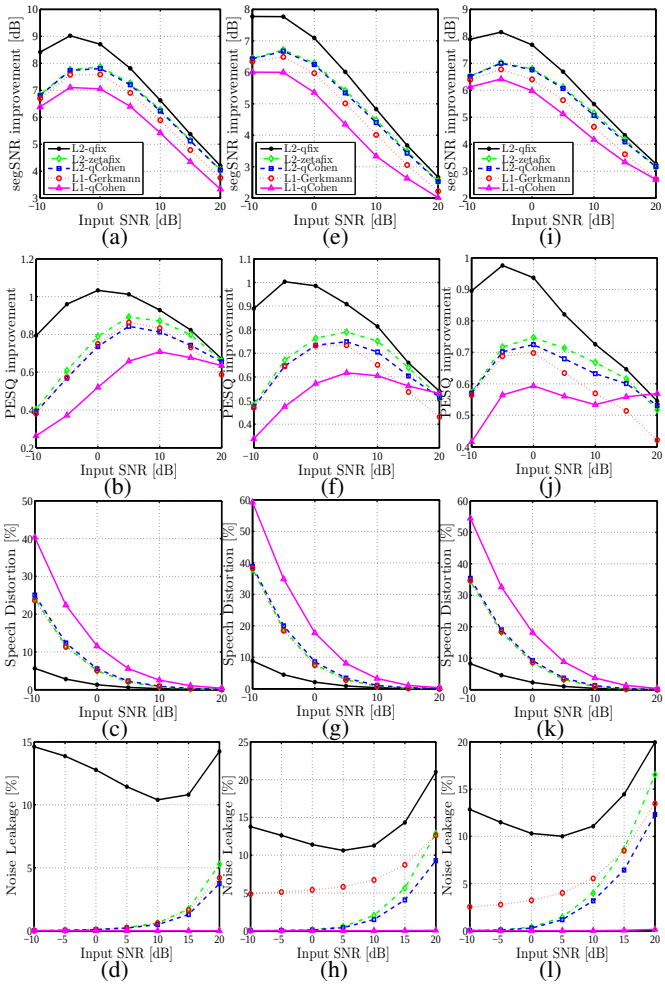

Fig. 2. The performance of different algorithms in terms of the improvements in seg. SNR and PESQ (with reference to unprocessed noisy input), SD and NL in various noise conditions: (a-d) white Gaussian, (e-h) Factory 1, and (i-l) babble noise.

7. REFERENCES

[1] 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.

[2] 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, July 2011.

[3] 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.

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

[5] 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.

[6] 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.

[7] H. Momeni, E. A. P. Habets, and H. R. Abutalebi, "Single-channel speech presence probability estimation using inter-frame and inter-band correlations," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 2927–2931.

[8] D. H. Tran Vu and R. Haeb-Umbach, "Using the turbo principle for exploiting temporal and spectral correlations in speech presence probability estimation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 863–867.

[9] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation exploiting temporal and spectral correlations using 2D-HMMs," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Sept 2013, pp. 1–5.

[10] C. M. Lee, J. W. Shin, and N. S. Kim, "Stereophonic acoustic echo suppression incorporating spectro-temporal correlations," *IEEE Signal Process. Lett.*, vol. 21, no. 3, pp. 316–320, March 2014.

[11] Y. G. Jin, J. W. Shin, and N. S. Kim, "Spectro-temporal filtering for multichannel speech enhancement in short-time Fourier transform domain," *IEEE Signal Process. Lett.*, vol. 21, no. 3, pp. 352–355, March 2014.

[12] 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, July 2008.

[13] B. Dashtbozorg and H. R. Abutalebi, "Adaptive MMSE speech spectral amplitude estimator under signal presence uncertainty," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aug 2009, pp. 209–212.

[14] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept 2012, pp. 1–4.

[15] M. Taseska and E.A.P Habets, "MMSE-based source extraction using position-based posterior probabilities," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 664–668.

[16] H. Momeni and H. R. Abutalebi, "Generalization of maximum *a posteriori* amplitude estimator under speech presence uncertainty for speech enhancement," *Journal of Circuits, Systems and Signal Processing*, vol. 33, no. 8, pp. 2565–2582, March 2014.

[17] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383 –1393, May 2012.

[18] M. Krawczyk-Becker, D. Fischer, and T. Gerkmann, "Utilizing spectro-temporal correlations for an improved speech presence probability based noise power estimation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 365–369.

[19] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[20] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation theory*, Fundamentals of Statistical Signal Processing. Prentice-Hall PTR, 1993.

[21] 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, July 2010.

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

[23] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan 2008.