Multiframe maximum a posteriori estimators for single-microphone speech enhancement

Raziyeh Ranjbaryan  |  Hamid Reza Abutalebi

Abstract
Multiframe maximum a posteriori (MAP) estimators are applied to a single-microphone noise reduction problem. Several attempts have been made to exploit the interframe correlation (IFC) between speech coefficients in the short-time Fourier transform domain. In a noise-reduction algorithm, all available information of recorded signals should be optimally utilized in the estimation process. Single-microphone multiframe minimum variance distortion-less response and single-microphone multiframe Wiener filters (MFWFs) have been presented in this approach. Incorporating the concept of IFC in the MAP estimator leads to multiframe MAP estimators in a single-microphone case. In each time-frequency unit, the current and a finite number of past noisy signals are utilized to develop the estimators. A complex factor is adopted to model the IFC between speech signals, which allows the application of multiframe MAP estimators. The noise reduction performance is compared for the proposed estimators with the joint MAP estimator (which ignores the correlation between successive frames) and benchmark MFWFs and speech-distortion weighted interframe Wiener filters for different input noise types. These evaluations verify that the proposed methods exhibit good performance.

1 | INTRODUCTION

It is barely possible to imagine life without sounds; people talk and communicate by voice. However, undesired background noise always exists and makes our hearing unpleasant, or in some cases, impossible. Speech enhancement has been a significant issue for researchers because of its several applications, such as hands-free, hearing aids, teleconferencing, human-to-human, and human-to-machine communication [1].

Because implementation of the fast Fourier transform is straightforward, noise-reduction algorithms are usually performed in the short-time Fourier transform (STFT) domain. In 1984, Ephraim and Malah introduced a minimum mean square error (MMSE) estimator to extract the short-time spectral amplitude (STSA) of clean speech from a noisy observed signal in the STFT domain. They assumed that successive speech frames are uncorrelated and considered noise and speech signals as statistical random variables with an independent Gaussian distribution in each time-frequency unit (TFU) [2]. They also derived an MMSE-STSA estimator under signal presence uncertainty and proposed a decision-directed approach to estimate a priori the signal-to-noise ratio (SNR). In addition, Ephraim and Malah extended their estimation to minimize the mean square error in the log-spectral domain, called MMSE-log STSA [3].

In [4], Wolfe and Godsill proposed simple alternatives to the MMSE-STSA rule concerning the maximum a posteriori (MAP) criterion. They developed MAP estimators to extract the phase and amplitude of a clean speech signal. These estimators assumed that successive speech frames are uncorrelated. It was also shown that the proposed joint MAP phase estimate is precisely equal to the noisy signal phase [4]. Both MMSE and MAP estimators were then generalized by other researchers, for example, by considering \(\beta\)-order MMSE criteria [5] or other probability distributions for speech signals [6, 7].

In [8], Lotter et al. extended the MMSE-STSA estimator (presented in [2]) to a multimicrophone case. First, they derived a multimicrophone estimation of MMSE-STSA, which requires knowledge of the direction of arrival (DOA) and confluent hypergeometric series. Second, using the magnitude of noisy observed signals, they introduced a direction-independent...
MAP estimator. Trawicki and Johnson in [9] presented an extension of the work of Lotter et al., wherein the optimal MMSE estimators for STSA, log-spectral amplitude, and spectral phase of clean speech signal were developed. The proposed spectral phase estimator considerably improves estimation accuracy but requires an initial time alignment.

All previously mentioned algorithms assume that successive speech frames are uncorrelated. However, it is well known that this assumption is not accurate. The low-pass nature of speech signal and its inherent coherence, along with the overlap procedure applied in the STFT, introduces high correlation, especially between consecutive frames. In the speech enhancement problem, we are primarily interested in utilizing all available information of the noisy signal. In recent years, the issue of interframe and interband correlations has received much attention and has been applied in many speech-related applications [10–16].

In [10], the authors considered interframe correlation (IFC) in the STFT domain for single-microphone noise reduction and extended two well-known beam-formers to the multiframe case. In each TFU, they utilized the current and a finite number of past noisy signals, enabling them to introduce a new model similar to the multicorephon case. Consequently, multiframe minimum variance distortionless response (MFMVDR) and multiframe Wiener filters (MFWFs) (which are usually employed in multimicrophone algorithms) were developed for the first time for single-microphone cases. It was shown that a significant noise reduction could be obtained using the MFMVDR and MWF if IFC can be accurately estimated.

Several researchers have probed the issue of IFC and its importance in speech enhancement. The sensitivity of the MFMVDR filter to estimation errors in the speech correlation matrix and corresponding speech IFC vector was investigated in [15]. In [12], the authors proposed blind maximum-likelihood (ML) and MAP estimators for the speech IFC vector. In addition, to set the trade-off between speech distortion and noise reduction, speech-distortion weighted interframe Wiener filters have been proposed in [16]. In [11], Momeni et al. proposed a single-microphone speech presence probability (SPP) estimator based on interframe and interband correlations. It was shown that detection accuracy is increased by taking these correlations into account. Applying the SPP presented in [13], they also proposed a conditional MMSE filter.

Although several contributions have focussed on beam-forming techniques, to the best of our knowledge, there has not been a detailed investigation on multiframe Bayesian estimators in the STFT domain that considers signal statistical properties. Inspired by the method presented in [10], we incorporate the concept of IFC in Bayesian estimators and introduce multiframe MAP estimators. Section 4 illustrates how the noise reduction performance will be affected by the use of MFMAP estimators compared with the single-microphone joint MAP estimator and the benchmark MFWTs. Finally, some concluding remarks are presented in Section 5.

2 | STATISTICAL PROPERTIES OF SIGNALS

Consider a single microphone that captures the clean speech signal $x(t)$ in a noisy field. We assume that the received signal is corrupted by uncorrelated additive noise $v(t)$. The noisy (observed) signal $y(t)$ at time index $t$ is given by $y(t) = x(t) + v(t)$. Using STFT, $Y_{m,k} = X_{m,k} + V_{m,k}$, where $Y_{m,k}$, $X_{m,k}$, and $V_{m,k}$ are the noisy signal, the clean speech, and the additive noise signals, respectively, with $m$ the frame index and $k$ the discrete frequency index. Let $Y_{m,k} = R_{m,k}e^{j\theta_{m,k}}$ and $X_{m,k} = A_{m,k}\alpha_{n,m}$ denote the spectral amplitude and phase of noisy and clean speech signals, respectively. For simplicity, we omit the discrete frequency index and only write it when referring to a specific TFU. It is assumed that the noise and clean speech signals are zero mean random processes. The main goal of this work is to estimate $A_{m}$ and $\alpha_{n,m}$.
It is assumed that the real and imaginary parts of clean speech signals are statistically independent Gaussian variables. The amplitude and phase of clean speech signals can be considered statistically independent and modeled as Rayleigh and uniform distributions, respectively, given by

\[
p(A_m) = \begin{cases} 
\frac{2A_m}{\sigma_x^2(m)} \exp\left(\frac{-A_m^2}{2\sigma_x^2(m)}\right) & 0 < A_m \\
0 & \text{else}
\end{cases}
\] (1)

and

\[
p(a_m) = \frac{1}{2\pi} - \pi < a_m < \pi.
\] (2)

Hence, their joint probability density function (pdf) is written as follows:

\[
p(A_m, a_m) = \frac{A_m}{\pi \sigma_x^2(m)} \exp\left(-\frac{A_m^2}{2\sigma_x^2(m)}\right),
\] (3)

where \(\sigma_x^2(m)\) denotes the variance in the clean speech signal at the time frame \(m\).

Under the assumption of Gaussian distribution for noise signals, the conditional pdf of \(Y_m\) given the amplitude and phase of the clean speech signal is given by

\[
p(Y_m | A_m, a_m) = \frac{1}{2\pi \sigma_n^2(m)} \exp\left(-\frac{|Y_m - A_me^{i\alpha}|^2}{2\sigma_n^2(m)}\right),
\] (4)

and \(\sigma_n^2(m)\) is the variance in the noise signal at the time frame \(m\).

3 | MULTIFRAME MAP ESTIMATORS

In recent years, IFC and its potential in speech enhancement have been reported in several research works. When a noise reduction algorithm is applied, it is important that all available information of recorded signals is optimally utilized. Thus, multiframe beam-forming algorithms by considering IFC have been presented in [10]. These algorithms aim to find optimal coefficients based on various criteria. They make no assumptions on the statistical properties of clean speech and noise signals. However, we take into account these properties to derive optimal MFMAP estimators in this work.

Like the mentioned multiframe beam-forming algorithms, in the \(m\)-th TFU, we utilize the current, and a finite number of past noisy signals in the same frequency index expressed as a noisy vector \(y_m\) such as

\[
y_m = [Y_m, Y_{m-1}, \ldots, Y_{m-L+1}]^T,
\] (5)

where the superscript \(T\) denotes transpose operation, and \(L\) is the length of the vector. It indicates how many frames have been used to model IFC. The vectors \(x_m\) and \(v_m\) can be defined similarly, such that \(y_m = x_m + v_m\). We also define \(r_m\) as the amplitude of noisy vector, that is,

\[
r_m = [R_m, R_{m-1}, \ldots, R_{m-L+1}]^T.
\] (6)

Assuming that the speech and noise signals are uncorrelated, the correlation matrix of the noisy signal is given by

\[
\Phi_{y_m} = \mathbb{E}\{y_m y_m^H\} = \Phi_{x_m} + \Phi_{v_m},
\] (7)

where \(\mathbb{E}\{\cdot\}\) denotes the expectation operator. \(\Phi_{x_m}\) and \(\Phi_{v_m}\) denote the correlation matrices of the clean speech and noise signals, respectively.

In practice, the correlation matrix of the noisy signal is recursively estimated as a linear combination of the correlation matrix at previous frames and the current information, that is,

\[
\hat{\Phi}_{y_m} = \lambda_y \hat{\Phi}_{y_{m-1}} + (1 - \lambda_y)y_m y_m^H.
\] (8)

Similarly in silent frames, the correlation matrix of noise is given by

\[
\Phi_{v_m} = \lambda_c \Phi_{v_{m-1}} + (1 - \lambda_c)v_m v_m^H,
\] (9)

where \(\lambda_y\) and \(\lambda_c\) denote forgetting factors. An estimate of the correlation matrix of the clean speech signal is given by

\[
\hat{\Phi}_{x_m} = \hat{\Phi}_{y_m} - \hat{\Phi}_{v_m}.
\] (10)

Because of the errors that occur in the estimation of correlation matrices, negative eigenvalues of \(\Phi_{x_m}\) should be set to zero to ensure that the resulting matrix is positive semidefinite.

Contrary to the multiframe beam-forming algorithms that make no assumptions on the IFC model, an appropriate model must derive the MFMAP estimators. This model represents the relation between the amplitude and phase of the clean speech in the complex STFT domain at the time frame \((m - i)\), \(X_{m-i}\), and the amplitude and phase of clean speech at the time frame \((m)\), \(X_{m}\). In this work, to address this issue, we propose a linear model for amplitudes, that is,

\[
A_{m-i} = C_{i,m} A_m.
\] (10)

A linear model simplifies the derivation of the estimators, as will be explained in the following results in quadratic expressions. We also model the phase difference between consecutive frames by multiplication to an exponential factor (considering the time shift property of fast Fourier transform). So, the IFC is totally modelled by a complex factor

\[
S_{i,m} = C_{i,m} e^{i\beta_{i,m}}
\] (where \(C_{i,m}\) affects the amplitude and \(\beta_{i,m}\)
indicates the phase shift). The correlation between the $X_{m-i}$ and $X_m$ can be modelled as

$$X_{m-i} = S_{m,i}X_m \Rightarrow A_{m-i}e^{j\alpha_{m-n}} = C_{i,m}A_m e^{j(\alpha_n - \beta_{i,m})}. \quad (11)$$

where $C$ is defined via

$$C_{i,m}^2 = \frac{\mathbb{E}\{X_{m-i}X_{m+i}\}}{\mathbb{E}\{X_mX_m\}} = \frac{\sigma^2(m-i)}{\sigma^2(m)}. \quad (12)$$

According to Equation (11), the current and past noisy signals can be written in terms of the clean speech amplitude and phase as follows:

$$\begin{align*}
Y_m &= C_{0,m}A_me^{j(\alpha_n - \beta_{i,m})} + V_m, \\
Y_{m-1} &= C_{1,m}A_me^{j(\alpha_n - \beta_{i,m})} + V_{m-1}, \\
\vdots \\
Y_{m-L+1} &= C_{L-1,m}A_me^{j(\alpha_n - \beta_{i,m})} + V_{m-L+1}.
\end{align*} \quad (13)$$

Obviously $C_{0,m} = 1$ and $\beta_{i,m} = 0$. Later, we explain how to estimate $\beta_{i,m}$'s.

Assuming that the noise signals at different frames are independent, the conditional pdf of $Y_m$ given the amplitude and phase of the clean speech signal at frame $m$ is given by

$$p(y_m | A_m, \alpha_m) = \prod_{i=0}^{L-1} p(Y_{m-i} | A_m, \alpha_m) = \prod_{i=0}^{L-1} \frac{1}{\pi\sigma^2(m-i)} \exp\left(-\sum_{i=0}^{L-1} \frac{|Y_{m-i} - C_{i,m}A_me^{j(\alpha_n - \beta_{i,m})}|^2}{\sigma^2(m-i)}\right), \quad (14)$$

where the variance in clean signal $\sigma^2(m-i)$ and the variance in noise signal at time frame $m-i$, that is, $\sigma^2(m-i)$, can be found as the diagonal elements of the speech and noise correlation matrices, respectively, that is,

$$\begin{align*}
\sigma^2(m-i) &= \Phi_{\alpha_n}(i+1, i+1) & i = 0, \ldots, L-1 \\
\sigma^2(m-i) &= \Phi_{\beta_n}(i+1, i+1) & i = 0, \ldots, L-1
\end{align*} \quad (15)$$

In the following, we employ the MAP criterion for single-microphone speech enhancement, resulting in multiframe MAP estimators. We first develop a joint MFMAP estimator for both the phase and amplitude of the clean speech signal, which can be interpreted as an extension of the single-microphone joint MAP estimator presented in [4] to the multiframe case. Next, we also derive two estimators for the amplitude only, while the phase is kept unchanged.

### 3.1 Joint multiframe MAP estimator for phase and amplitude

Here, we propose an algorithm that estimates both spectral phase and amplitude by maximizing the posterior distribution of $A_m$ and $\alpha_m$ given the noisy observed vector $y_m$ as

$$\hat{A}_m, \hat{\alpha}_m = \arg\max_{A_m, \alpha_m} p(A_m, \alpha_m | y_m) = \arg\max_{A_m, \alpha_m} \frac{p(y_m | A_m, \alpha_m)p(A_m, \alpha_m)}{p(y_m)}. \quad (16)$$

Because the denominator is independent of phase and amplitude, it is sufficient to maximize only the numerator, as

$$\hat{A}_m, \hat{\alpha}_m = \arg\max_{A_m, \alpha_m} p(y_m | A_m, \alpha_m) \quad (17)$$

To find the optimal JMFMAP estimator, we combine Equations (14) and (3), and obtain the numerator of Equation (16) as

$$p(y_m | A_m, \alpha_m) = \prod_{i=0}^{L-1} \frac{1}{\pi\sigma^2(m-i)} \exp\left(-\sum_{i=0}^{L-1} \frac{|Y_{m-i} - C_{i,m}A_me^{j(\alpha_n - \beta_{i,m})}|^2}{\sigma^2(m-i)}\right) \frac{A_m}{\sigma^2(m)} \exp\left(-\frac{A^2_m}{\sigma^2(m)}\right). \quad (18)$$

Because $\ln(\cdot)$ is a monotonically increasing function, the maximization problem in Equation (17) is equivalent to maximizing the natural logarithm of it. Ignoring the terms that do not affect the optimization, we proceed with

$$\begin{align*}
\ln p(y_m | A_m, \alpha_m) &= -\sum_{i=0}^{L-1} \frac{|Y_{m-i} - C_{i,m}A_me^{j(\alpha_n - \beta_{i,m})}|^2}{\sigma^2(m-i)} \\
&\quad + \ln A_m - \frac{A^2_m}{\sigma^2(m)} \quad (19)
\end{align*}$$

### 3.1.1 Joint multiframe MAP estimator to extract clean signal phase

Considering the important role of phase processing, we firstly focus on exploiting the phase of the clean signal. This is carried out by maximizing Equation (19) with respect to $\alpha_m$. To this end, by differentiating Equation (19) with respect to $\alpha_m$ and setting it to zero, we obtain
\[ \frac{\partial}{\partial \alpha_m} \ln p(y_m, A_m, \alpha_m) = \]

\[ \sum_{i=0}^{L-1} \left[ Y_{m-i} - C_{i,m} A_m e^{j(\alpha_m - \beta_m)} \right] \frac{j C_{i,m} A_m e^{-j(\alpha_m - \beta_m)}}{\sigma_v^2(m-i)} \]

\[ + \sum_{i=0}^{L-1} \left[ Y^*_{m-i} - C_{i,m} A_m e^{-j(\alpha_m - \beta_m)} \right] \frac{-j C_{i,m} A_m e^{-j(\alpha_m - \beta_m)}}{\sigma_v^2(m-i)} = 0. \]

Substituting \( Y_{m-i} = R_{m-i} e^{j\theta_{m-i}} \) and further simplifications, the following closed-form equation is obtained:

\[ \frac{1}{\sigma_v^2(m-i)} \sum_{i=0}^{L-1} C_{i,m} R_{m-i} \sin(\theta_{m-i} + \beta_i,m - \alpha_m) = 0, \]

where \( n_{i,m} = \frac{C_{i,m} R_{m-i}}{\sigma_v^2(m-i)} \) and \( \theta_{i,m} = \theta_{m-i} + \beta_i,m \).

To find the optimal JMFMAP estimator of phase, we use the trigonometry relation [20], indicating that the sum of a finite number of sine terms with different phases and amplitudes can be represented as one single term,

\[ \sum_i d_i \sin(\eta_i - \omega) = d \sin(\eta - \omega), \]

where the resulting amplitude and phase are given by

\[ \begin{cases} 
  d^2 = \sum_{i,j} d_i d_j \cos(\eta_i - \eta_j), \\
  \eta = \arctan \left( \frac{\sum_i d_i \sin(\eta_i)}{\sum_i d_i \cos(\eta_i)} \right). 
\end{cases} \]

Consequently, the optimal estimator of the phase is obtained from Equation (21) as

\[ \alpha_m = \arctan \left( \frac{\sum_{i=0}^{L-1} n_{i,m} \sin(\theta_{m-i} + \beta_i,m)}{\sum_{i=0}^{L-1} n_{i,m} \cos(\theta_{m-i} + \beta_i,m)} \right). \]

In the case of \( L = 1 \) (i.e., exploiting the current frame only), it is observed that Equation (24) becomes equal to the noisy phase, which is also the joint MAP estimation of phase in the single-microphone case as presented in [4].

To provide the optimal estimation of phase, it is required to compute \( n_{i,m} \) and \( \beta_i,m \). To this end, we considered a priori and a posteriori SNRs, respectively,

\[ \zeta_{m-i} = \frac{\sigma_v^2(m-i)}{\sigma_v^2(m-i)} \quad \gamma_{m-i} = \frac{R_{m-i}^2}{\sigma_v^2(m-i)} \]

\( R_{m-i}^2 \) is directly computed according to the amplitude of the noisy (observed) signals. Using above definitions and Equation (12), \( n_{i,m} \) is written in terms of the a priori and a posteriori SNRs as follows:

\[ \frac{n_{i,m}}{\sigma^2_v(m-i)} = \frac{\sigma_v(m-i)}{\sigma_v^2(m-i)} \frac{R_{m-i}}{\sigma_v^2(m-i)} = \frac{1}{\sigma_v(m)} \sqrt{\zeta_{m-i} y_{m-i}}. \]

To estimate \( \beta_i,m \), we consider the estimation of signal parameters via rotational invariance techniques (ESPRIT) algorithm [21], which is usually employed for DOA estimation using microphone arrays. When the array comprises two subarrays, and only the phases are of difference, the ESPRIT algorithm is convenient for exploiting the phase difference. A brief introduction of the ESPRIT algorithm has been provided in the Appendix.

Inspired by the idea behind ESPRIT, we present a method to estimate \( \beta_i,m \); for example, to estimate \( \beta_{2,m} \) (the phase difference between the clean signal at the \( m \)-th time frame, \( X_m \), and the clean signal at the \( (m-2) \)-th time frame, \( X_{m-2} \)), only the noisy observed signals are applied such as

\[ \begin{align*}
  Y_m &= e^{j\theta_m} A_m + V_m, \\
  Y_{m-2} &= C_{2,m} e^{j\theta_{m-2}} A_m + V_{m-2}.
\end{align*} \]

It should be noted that the noisy observed signal at each TFU is considered as one subarray in our problem. The model presented in Equation (27) makes it convenient to estimate \( \beta_{2,m} \) [21]. To this end, we first need to define a new vector containing \( Y_m \) and \( Y_{m-2} \), that is, \( \tilde{z}_{2,m} = [Y_m, Y_{m-2}]^T \), and compute its correlation matrix, using a recursive approach similar to Equation (7), as

\[ \Phi_{\tilde{z}_{2,m}} = \lambda \Phi_{\tilde{z}_{2,m-1}} + (1 - \lambda) \tilde{z}_{2,m} \tilde{z}_{2,m}^H. \]

The ESPRIT algorithm is applied to \( \Phi_{\tilde{z}_{2,m}} \) to compute \( \beta_{2,m} \). This procedure is repeated until all \( \beta_S \) have been calculated.

3.1.2 Joint multiframe MAP estimator to extract the amplitude of clean signal

In this subsection, the proposed JMFMAP estimator aims to find the optimal spectral amplitude. To this end, after computing the spectral phase based on Equation (24), we maximize the posterior distribution Equation (19) with respect to \( A_m \). Differentiating with regard to \( A_m \) and setting the result equal to zero yields
By substituting $Y_{m-i} = R_{m-i}e^{\theta_{m-i}}$, the following is obtained:

$$2 \sum_{i=0}^{L-1} C_{i,m} R_{m-i} \cos(\theta_{m-i} + \beta_{i,m} - \alpha_m) \sigma_z^2(m-i) - 2A_m \sum_{i=0}^{L-1} \frac{C_{i,m}^2}{\sigma_z^2(m-i)} + \frac{1}{A_m} - \frac{2A_m}{\sigma_z^2(m)} = 0, \quad \text{(30)}$$

which represents a quadratic function in terms of $A_m$. Using Equation (26) and

$$\sum_{i=0}^{L-1} \frac{C_{i,m}^2}{\sigma_z^2(m-i)} + \frac{1}{A_m} - \frac{2A_m}{\sigma_z^2(m)} = 0, \quad \text{(31)}$$

we can rewrite Equation (30) in terms of a priori and a posteriori SNRs,

$$A_m^2 \left( 2 + 2 \sum_{i=0}^{L-1} \xi_{m-i} \right) + A_m \left( -2\sigma_x(m) \sum_{i=0}^{L-1} \sqrt{\xi_{m-i}\lambda_{m-i}} \cos(\theta_{m-i} + \beta_{i,m} - \alpha_m) \right) - \sigma_z^2(m) = 0. \quad \text{(32)}$$

The JMFMAP estimator of amplitude is computed by applying a gain factor to the noisy amplitude (i.e. $A_m = G_m R_m$). The gain $G_m$ is achieved as

$$G_m = \frac{A_m}{R_m} = \frac{\sqrt{\gamma_m}}{2 + 2 \sum_{i=0}^{L-1} \sqrt{\xi_{m-i}\lambda_{m-i}} \cos(\phi_i)} \Re \left\{ \sum_{i=0}^{L-1} \sqrt{\xi_{m-i}\lambda_{m-i}} \cos(\phi_i) \right\} + \sqrt{\left( \sum_{i=0}^{L-1} \sqrt{\xi_{m-i}\lambda_{m-i}} \cos(\phi_i) \right)^2 + 2 \left( 1 + \sum_{i=0}^{L-1} \xi_{m-i} \right)} \right), \quad \text{(33)}$$

where $\phi_i = \theta_{m-i} + \beta_{i,m} - \alpha_m$. Again, it can be seen that in the case of $L = 1$, Equation (33) will be simplified to the joint MAP estimation of amplitude in the single-microphone case as presented in [4].

Briefly, the proposed JMFMAP estimator operates as follows:

1. In each TFU, define the noisy signal vector as Equation (5).
2. Compute the correlation matrices of Equations (7)–(9).
3. Compute the variances of clean and noise signals based on Equation (15).
4. Compute a priori and a posteriori SNRs according to Equation (25).
5. Compute $n_{i,m}$ as stated by Equation (26) and $\beta_{i,m}$ using ESPRIT.
6. Compute the phase and amplitude according to Equations (24) and (33), respectively.

As described before, we presented the JMFMAP estimators for the phase and amplitude of the clean signal. It was observed that the phase difference between consecutive frames is required. For this purpose, we employed an ESPRIT-like algorithm. We propose two other estimators that only extract the amplitude while keeping the phase unprocessed.

### 3.2 | Multiframe MAP estimators of amplitude

In this section, we present two different MFMAP estimators of amplitude (or briefly, MFMAPEA-Ss) that respectively utilize the noisy signal vector ($y_m$), or only its amplitude ($a_m$). The first estimator can be considered an extension of the single-microphone MAP estimator of amplitude (by Wolfe and Godsill in [4]) to the multiframe case. The second one uses the results of the multimicrophone MAP estimator by Lotter et al. in [8].

#### 3.2.1 | Multiframe MAP estimator of amplitude given the noisy signal vector

Assuming the noisy signal vector is available, we propose an estimator that extracts the clean signal’s amplitude and keeps its phase unchanged. The MFMAPEA-Sig estimator only considers the spectral amplitude and aims to maximize the posterior distribution of $A_m$ given the noisy vector, that is,

$$\hat{A}_m = \arg\max_{A_m} p(A_m | y_m) = \arg\max_{A_m} \frac{p(y_m, A_m)}{p(y_m)}, \quad \text{(34)}$$

Similar to what was described in III-A, it is sufficient to maximize only the numerator in Equation (34).

In this section, our main objective is to estimate only the amplitude. Using Equation (14), the following joint pdf is obtained as
\[
\begin{align*}
    p(y_m, A_m) &= \int_0^{2\pi} p(y_m | A_m, a_m) p(A_m, a_m) \, da_m = \\
    &\frac{1}{\pi} \prod_{i=0}^{L-1} \frac{A_m}{\pi \sigma_c^2(m-i)} \exp \left( -\frac{A_m^2}{\sigma_c^2(m)} \right) \exp \left( -\sum_{i=0}^{L-1} \frac{|Y_{m-i} - C_{i,m} A_m e^{(a_m-\beta_m)}|^2}{\sigma_c^2(m-i)} \right) \, da_m. \\
    \text{The integral is simplified as} \ [9]: \quad &
\end{align*}
\]

\[
\begin{align*}
    \int_0^{2\pi} \exp \left( -\sum_{i=0}^{L-1} \frac{|Y_{m-i} - C_{i,m} A_m e^{(a_m-\beta_m)}|^2}{\sigma_c^2(m-i)} \right) \, da_m = \\
    \exp \left( -\sum_{i=0}^{L-1} \frac{|Y_{m-i}|^2 + (C_{i,m} A_m)^2}{\sigma_c^2(m-i)} \right) \\
    \int_0^{2\pi} \exp(a \cos(a_m) + b \sin(a_m)) \, da_m = \\
    2\pi I_0 \left( 2A_m \sum_{i=0}^{L-1} \frac{C_{i,m} Y_{m-i} e^{\beta_m}}{\sigma_c^2(m-i)} \right), \\
\end{align*}
\]

where \( a = \sum_{i=0}^{L-1} 2C_{i,m} A_m \Re\{Y_{m-i} e^{\beta_m}\} \), \( b = \sum_{i=0}^{L-1} 2C_{i,m} A_m \Im\{Y_{m-i} e^{\beta_m}\} \).

In addition, the summation of two cosine terms is given by [9].

\[
\begin{align*}
    a \cos(a_m) + b \sin(a_m) &= \sqrt{a^2 + b^2} \cos(a_m - \arctan(b/a)), \\
\end{align*}
\]

where \( \sqrt{a^2 + b^2} = 2A_m \left| \sum_{i=0}^{L-1} C_{i,m} Y_{m-i} e^{\beta_m} \right| \). Next, we obtain

\[
\begin{align*}
    \int_0^{2\pi} \exp(a \cos(a_m) + b \sin(a_m)) \, da_m = \\
    2\pi I_0 \left( 2A_m \sum_{i=0}^{L-1} \frac{C_{i,m} Y_{m-i} e^{\beta_m}}{\sigma_c^2(m-i)} \right), \\
\end{align*}
\]

where \( I_0(x) \) is the modified Bessel function of the first kind of the 0th order. Finally, the joint pdf of \( y_m \) and \( A_m \) is given by

\[
\begin{align*}
    p(y_m, A_m) &= \prod_{i=0}^{L-1} \frac{1}{\pi \sigma_c^2(m-i)} \frac{A_m}{\pi \sigma_c^2(m)} \exp \left( -\frac{A_m^2}{\sigma_c^2(m)} \right) \\
    &\exp \left( -\sum_{i=0}^{L-1} \frac{|Y_{m-i}|^2 + (C_{i,m} A_m)^2}{\sigma_c^2(m-i)} \right) \\
    &2\pi I_0 \left( 2A_m \sum_{i=0}^{L-1} \frac{C_{i,m} Y_{m-i} e^{\beta_m}}{\sigma_c^2(m-i)} \right). \\
\end{align*}
\]

Ignoring the terms that do not affect the optimization, we proceed with

\[
\begin{align*}
    \ln p(y_m, A_m) &= \ln(A_m) - \frac{A_m^2}{\sigma_c^2(m)} - \sum_{i=0}^{L-1} \frac{|Y_{m-i}|^2 + (C_{i,m} A_m)^2}{\sigma_c^2(m-i)} \\
    &\ln \left( 2\pi I_0 \left( 2A_m \sum_{i=0}^{L-1} \frac{C_{i,m} Y_{m-i} e^{\beta_m}}{\sigma_c^2(m-i)} \right) \right). \\
\end{align*}
\]

Applying \( I_0(x) \approx \frac{1}{\sqrt{2\pi x}} e^x \) for large \( x \), Equation (41) is simplified to

\[
\begin{align*}
    \ln p(y_m, A_m) &= \ln(A_m) - \frac{A_m^2}{\sigma_c^2(m)} - \sum_{i=0}^{L-1} \frac{|Y_{m-i}|^2 + (C_{i,m} A_m)^2}{\sigma_c^2(m-i)} \\
    &- \frac{1}{2} \ln \left( \frac{4\pi A_m \left| \sum_{i=0}^{L-1} C_{i,m} Y_{m-i} e^{\beta_m} \right|}{\sigma_c^2(m-i)} \right) \\
    &+ 2A_m \sum_{i=0}^{L-1} \frac{C_{i,m} Y_{m-i} e^{\beta_m}}{\sigma_c^2(m-i)}. \\
\end{align*}
\]

The spectral amplitude that maximizes the posterior distribution given the noisy vector is computed by differentiating Equation (42) with respect to \( A_m \) and setting the result equal to zero as

\[
\begin{align*}
    A_m^2 \left( -\frac{1}{\sigma_c^2(m)} - \sum_{i=0}^{L-1} \frac{C_{i,m}^2}{\sigma_c^2(m-i)} \right) \\
    + A_m \left| \sum_{i=0}^{L-1} C_{i,m} Y_{m-i} e^{\beta_m} \right| + \frac{1}{4} = 0. \\
\end{align*}
\]

This represents a quadratic function in terms of \( A_m \). Using Equations (26) and (31), the quadratic function can be written as follows:

\[
\begin{align*}
    A_m^2 \left( 1 + \sum_{i=0}^{L-1} \zeta_{m-i} \right) + \\
    A_m \left( -\sigma_c(m) \sqrt{\sum_{i=0}^{L-1} \zeta_{m-i} e^{i(\theta_{m-i} + \beta_i)}} \right) - \frac{\sigma_c^2(m)}{4} = 0. \\
\end{align*}
\]

The MFMAPEA-Sig estimator is computed by applying a gain factor to the noisy amplitude (i.e. \( A_m = G_m R_m \)). The gain \( G_m \) is achieved as
\[
G_m = \frac{A_m}{R_m} = \frac{\sqrt{y_m}}{2 + 2 \sum_{i=1}^{L-1} \xi_{m-i}} \\
\text{Re} \left\{ \frac{1}{2} \left( \sum_{i=0}^{L-1} \sqrt{\xi_{m-i} y_{m-i}} e^{(i \theta_{m-i} + \beta_m)} \right) \right\} + \sqrt{\left( \sum_{i=0}^{L-1} \sqrt{\xi_{m-i} y_{m-i}} e^{(i \theta_{m-i} + \beta_m)} \right)^2 + \left( 1 + \sum_{i=0}^{L-1} \xi_{m-i} \right)}\right\}.
\]

We also observe that in the special case of \( L = 1 \), Equation (45) will be simplified to the MAP estimation of amplitude in the single-microphone case as presented in [4].

Briefly, the proposed MFMAPEA-Sig operates as follows:

1. In each TFU, define the noisy signal vector as Equation (5).
2. Compute the correlation matrices of Equations (7)–(9).
3. Compute the variance in clean and noise signals based on Equation (15).
4. Compute a priori and a posteriori SNRs according to Equation (25).
5. Compute \( \beta_{i,m} \) using ESPRIT.
6. Compute the amplitude as stated by Equation (45).
7. Keep the phase unchanged.

Again, it is seen that we need to calculate the phase difference between consecutive frames, which in turn requires an additional procedure and extra computational load. To overcome this problem and provide a less complex estimator, we propose another MFMAP estimator for amplitude in the next subsection, which does not require any knowledge of the parameter \( \beta_{i,m} \).

### Multiframe MAP estimator of amplitude given the amplitude of the noisy vector

In [8], Lotter et al. proposed a direction-independent MAP estimator for amplitude in the multimicrophone case given the noisy signal's amplitude. Inspired by their work, we propose our second MFMAPEA, which aims to maximize the posterior distribution of \( A_m \) given the amplitude of the noisy signal vector (instead of the whole noisy signal vector).

We propose the MFMAPEA-Amp algorithm that can be considered the multichannel MAP estimator (presented in [8]) to the multiframe case.

The optimization problem can be written as follows:

\[
\hat{A}_m = \arg \max_{A_m} p(A_m | r_m) = \arg \max_{A_m} \frac{p(A_m, r_m)}{p(r_m)}.
\]

Because the steps to obtain the optimal gain are similar to Lotter et al. work, we refer the readers to [8] for more details. In our proposed MFMAPEA-Amp, instead of the amplitude of noisy signals at different microphones, we utilize the amplitude of the current and a finite number of past noisy signals.

The proposed MFMAPEA-Amp is practically computed by applying a gain factor to the noisy amplitude (i.e., \( A_m = G_m R_m \)). The gain \( G_m \) is achieved as

\[
G_m = \frac{A_m}{R_m} = \frac{\sqrt{y_m}}{2 + 2 \sum_{i=1}^{L-1} \xi_{m-i}} \text{Re} \left\{ \frac{1}{2} \left( \sum_{i=0}^{L-1} \sqrt{\xi_{m-i} y_{m-i}} e^{(i \theta_{m-i} + \beta_m)} \right) \right\} + \sqrt{\left( \sum_{i=0}^{L-1} \sqrt{\xi_{m-i} y_{m-i}} e^{(i \theta_{m-i} + \beta_m)} \right)^2 + \left( 2 - L \right) \left( 1 + \sum_{i=0}^{L-1} \xi_{m-i} \right)}\right\}.
\]

Briefly, the proposed MFMAPEA-Amp operates as follows:

1. In each TFU, define the noisy signal vector as Equation (5).
2. Compute the correlation matrices of Equations (7)–(9).
3. Compute the variance in clean and noise signals based on Equation (15).
4. Compute a priori and a posteriori SNRs according to Equation (25).
5. Compute the amplitude as stated by Equation (47).
6. Keep the phase unchanged.

### SIMULATION RESULTS

In this section, to demonstrate the important role of interframe correlation, we report our comparative evaluations on the noise reduction performance of the proposed MFMAP estimators and three baseline methods: the joint MAP (JMAP) estimator of both phase and amplitude presented by Wolfe and Godsill in [4], where the correlation between successive frames is ignored (two first proposed estimators, JMFMAP and MFMAPEA, are the generalization of the JMAP to the multiframe case), and the benchmark MFWFs [10] and speech-distortion weighted interframe Wiener filter (SDW-IFWF) [16], where the current and a finite number of past noisy signals are utilized to develop MFWFs. The trade-off parameter in SDW-IFWF was set to \( \mu = 0.9 \), which seemed to give the best performance. To demonstrate the potential of phase processing for noise reduction, we also use the oracle value of \( \beta \) and compare the JMFMAP estimator’s performance when the accurate estimation of \( \beta \) is provided (named JMFMAP-\( \beta \)). \( \beta \) indicates the spectral speech phase difference between consecutive frames.

The performance of the considered estimators is compared in terms of four objective measures, namely perceptual evaluation of speech quality (PESQ) [22], short-time objective intelligibility (STOI) [23], segmental SNR (SegSNR) [24] improvement between the enhanced signal and the noisy reference signal, and finally, log-spectral distortion (LSD) [25]. The clean speech signal has been used as the reference signal.
The PESQ value is used as a measure for speech quality. STOI is a measure to evaluate the intelligibility of speech. SegSNR considers both noise reduction and speech distortion, and LSD value is defined as a measure for speech distortion; the lower the LSD, the lower the speech distortion. To compute the SegSNR values, an ideal voice activity detection (VAD) is applied to distinguish the speech frames.

We used samples of eight male and eight female speakers from the TIMIT database [26] during the first experiment as the clean speech samples. TIMIT is a well-known data set of clean speech widely used in artificially production of noisy speech by adding to the different noise signals. The presented evaluation results are the average of these 16 samples. Additive noises corrupt the microphone signals at input full-band SNRs ranging from −10 to 15 dB. Signals are recorded at sampling frequency \( f_s = 16 \) kHz. The STFT processing is implemented using NFFT = 256 with 75%-overlapping frames and using Hamming analysis window.

To apply JMAP estimation, it is required to compute \( \hat{\zeta}_m \) and \( \hat{\gamma}_m \) given by

\[
\hat{\zeta}_m = \frac{\hat{\sigma}^2(m)}{\sigma_v^2(m)}, \quad \hat{\gamma}_m = \frac{R^2_m}{\sigma_v^2(m)}. \tag{48}
\]

In practice, a decision-directed approach [2] is applied to estimate \( \zeta_m \) as

\[
\hat{\zeta}_m = a_c \frac{A_{m-1}}{\sigma_v^2(m)} + (1 - a_c) \max[0, \gamma_m - 1], \tag{49}
\]

where \( a_c \) is a forgetting factor. In addition, \( R^2_m \) is directly computed using the amplitude of noisy signal at time frame \( m \).

In the following, we evaluate the noise reduction performance based on two scenarios: (1) when the correlation matrix of noise is updated based on the noise signals corresponding to Equation (8) to avoid the effect of the estimation error of noise variance; and (2) where only the noisy signals are available.

### 4.1 Correlation matrix of noise updated based on noise signals

First, we discovered that the best performance is experimentally achieved by choosing \( \lambda_y = \lambda_v = 0.8 \) in Equations (7) and (8), respectively. In addition, we set \( a_c = 0.98 \) in Equation (49) as presented in [2].

For white Gaussian noise at input full-band SNR = 5 dB, Figure 1 depicts the performance of the proposed MFMAP estimators in terms of PESQ, STOI, and SegSNR improvement and LSD for several numbers of successive frames \( L \).

We observe from Figure 1(a) that the PESQ improvement provided by selecting \( L > 1 \) is better than \( L = 1 \) for all MFMAP estimators. It also depicts that \( L = 5 \) results in the best PESQ performance for all MFMAP estimators. This is mainly because for \( L = 1 \), regardless of interframe correlation, each TFU has been processed independently. However, for speech enhancement in single-microphone scenarios, it is essential to utilize all available information of a recorded signal. Indeed, the harmonic and quasi-periodic nature of vowels and some voiced speech leads to slowly varying amplitudes, introducing a high interframe correlation between consequence frames.

In terms of STOI, Figure 1(b) shows that all MFMAP estimators take the largest value by selecting \( L = 4 \). As expected, JMFMAP-\( \beta \) and JMFMAP, which estimate the clean speech phase, outperform the others. This result can be supported by [17, 27], verifying the perceptual importance of phase. In [17, 27], it was shown that the enhanced spectral speech phase has an important role in improving speech quality and intelligibility.

In terms of SegSNR, Figure 1(c) indicates that the SegSNR value for \( L > 1 \) is improved compared with \( L = 1 \) for JMFMAP-\( \beta \) and MFMAPEA-Amp. For JMFMAP and MFMAPEA-Sig estimators, although the SegSNR improvement for \( L = 2 \) is a bit more than \( L = 1 \), the change is not so sensible. This can be justified by considering the errors in the \( \beta \)-estimation procedure that reduce JMFMAP and MFMAPEA-Sig performance. However, the MFMAPEA-Amp estimator, which makes no assumption and consequently does not need to estimate this parameter, represents good performance. These results emphasize that the impact of \( \beta \)-parameter is crucial.

Figure 1(d) illustrates that the LSD value of JMFMAP-\( \beta \), JMFMAP, and MFMAPEA-Sig for \( L = 2 \) is slightly less than \( L = 1 \). The best performance for MFMAPEA-Amp is obtained by selecting \( L = 5 \). Considering the overlap of the speech signal and the noise in both time and frequency domains, part of the speech signal is also eliminated when we reduce the noise. Therefore, noise reduction usually comes at the price of signal distortion.

Note that by increasing the number of \( L \), the proposed linear model in Equation (10) may be inaccurate and introduce some errors. We select the number of successive frames as \( L = 4 \) in the following simulations. In our simulation, the STFT processing was implemented using NFFT = 256 at the sampling frequency \( f_s = 16 \) kHz, yielding 32 ms temporal correlation.

Table 1 depicts the performance of the proposed MFMAP estimators, the joint MAP estimator, MFWF, and SDW-IFWF considering additive white Gaussian noise at input full-band SNRs ranging from −10 to 15 dB. To provide a fair comparison when selecting \( L = 4 \) for the multiframe processing, we apply 32 ms frame length for JMAP, which only considers the information of the current frame (\( L = 1 \)). We have highlighted the best performance for each SNR in bold.

It can be seen that the MFWF and SDW-IFWF provide greater PESQ improvement. As expected, among the proposed MFMAP estimators, JMFMAP-\( \beta \) delivers the best performance for all input full-band SNRs. Next is MFMAPEA-Amp, which assumes nothing about parameter \( \beta_{i,m} \). It should be noted that although the JMFMAP estimator extracts both spectral phase and amplitude, the estimation errors of \( \beta_{i,m} \)...
impair the performance. To estimate $\beta_{im}$, we apply the ESPRIT algorithm. The ESPRIT algorithm is sensitive to the number of sensors in each subarray. Considering the single-microphone nature of our proposed method, each subarray consists of only one observation; this may introduce some error in computing $\beta_{im}$. This leads to the conclusion that further investigation is required to improve the robustness of JMFMAP and MFMAPEA-Sig against $\beta_{im}$ estimation or improve the estimation accuracy.

Concerning STOI, the MFWF and SDW-IFWF outperform the rest estimators in low- and mid-SNRs. It is also seen that multiframe MAP estimators deliver higher STOI performance than JMAP, especially in low SNRs. Comparing the proposed MFMAP estimators, as expected, JMFMAP-$\beta$ results in the best performance. Moreover, the performance of the JMFMAP estimator and the MFMAPEA-Sig is slightly better than the MFMAPEA-Amp.

Regarding SegSNR, Table 1 indicates that the MFWF, JMFMAP-$\beta$, and the joint MAP estimator perform the best in low, mid, and high SNRs, respectively. Moreover, the MFMAPEA-Amp obtains greater SegSNR improvement than JMFMAP-$\beta$ in low SNRs.

Finally, we investigate LSD in the presence of white Gaussian noise. Following the SDW-IFWF and MFWF, the MFMAPEA-Amp produces the best performance among the multiframe estimators. Furthermore, the JMAP estimator delivers the best performance in high SNRs.

It should be noted that there are two parallel comparisons; compared with the JMAP estimator, which ignores the effect of correlation between successive frames, by using the proposed MFMAP estimators, a significant speech quality improvement can be obtained. These results demonstrate that it is possible to improve some existing single-frame Bayesian methods (in this case, MAP
Table 1: Performance of the joint MAP estimator, proposed MFMAP estimators, MFWF, and SDW-IFWF in terms of PESQ, STOI, SegSNR improvement, and LSD in the case of white Gaussian noise for several input full-band SNRs when the exact VAD is available.

| Methods\Input SNR (dB) | PESQ Improvement | STOI Improvement |
|------------------------|------------------|------------------|
|                         | -10              | -5               | 0     | 5     | 10    | 15    | 5     | 10    |
| JMAP                   | 0.63             | 0.77             | 0.85  | 0.81  | 0.70  | 0.59  | JMAP  | 0.06  | 0.08  | 0.07  | 0.03  | 0.011 | 0.00  |
| JMFMAP                 | 1.31             | 1.28             | 1.17  | 1.03  | 0.87  | 0.67  | 1.31  | 0.25  | 0.17  | 0.10  | 0.043 | 0.007 |
| JMFMAP-beta            | 1.50             | 1.48             | 1.38  | 1.26  | 1.10  | 0.94  | 1.50  | 0.32  | 0.26  | 0.18  | 0.11  | 0.053 | 0.016 |
| MFMAPEA-Sig            | 1.25             | 1.24             | 1.16  | 1.04  | 0.90  | 0.75  | 1.25  | 0.30  | 0.24  | 0.17  | 0.09  | 0.042 | 0.011 |
| MFMAPEA-Amp            | 1.44             | 1.44             | 1.36  | 1.21  | 1.04  | 0.84  | 1.44  | 0.28  | 0.22  | 0.16  | 0.09  | 0.042 | 0.012 |
| MFWF                   | 1.84             | 1.79             | 1.65  | 1.44  | 1.17  | 0.91  | 1.84  | 0.34  | 0.27  | 0.18  | 0.11  | 0.049 | 0.011 |
| SDW-IFWF               | 1.80             | 1.81             | 1.48  | 1.38  | 1.15  | 0.86  | 1.80  | 0.34  | 0.26  | 0.19  | 0.10  | 0.048 | 0.009 |

| Methods\Input SNR (dB) | SegSNR Improvement |
|------------------------|---------------------|
|                         | JMAP                |
|                         | 15.54               | 13.64             | 11.45 | 9.25  | 7.20  | 5.35  | JMAP  | 7.71  | 5.61  | 4.08  | 3.13  | 2.47  | 1.83  |
|                         | JMFMAP              | 15.87             | 13.42 | 10.71 | 7.88  | 4.68  | 1.017 | JMFMAP| 6.69  | 5.31  | 4.25  | 3.35  | 2.66  | 2.19  |
|                         | JMFMAP-Beta         | 16.46             | 14.43 | 12.36 | 10.31 | 8.03  | 5.35  | JMFMAP-beta | 6.58  | 5.21  | 4.15  | 3.26  | 2.59  | 2.08  |
|                         | MFMAPEA-Sig         | 15.84             | 13.47 | 11.10 | 8.85  | 6.54  | 3.90  | MFMAPEA-Sig | 5.48  | 4.45  | 3.72  | 3.08  | 2.56  | 2.16  |
|                         | MFMAPEA-Amp         | 18.35             | 15.31 | 12.37 | 9.71  | 7.10  | 4.32  | MFMAPEA-Amp | 3.67  | 3.41  | 3.22  | 2.91  | 2.49  | 2.05  |
|                         | MFWF                | 20.67             | 17.14 | 13.54 | 9.94  | 6.16  | 2.12  | MFWF | 3.32  | 3.05  | 2.82  | 2.54  | 2.27  | 2.03  |
|                         | SDW-IFWF            | 20.51             | 17.06 | 13.39 | 9.80  | 6.05  | 2.01  | SDW-IFWF | 3.31  | 3.03  | 2.81  | 2.55  | 2.27  | 2.05  |

Note: Bolded items show the best performance in each SNR.

Abbreviations: JMAP, joint MAP; JMFMAP, joint multiframe MAP; LSD, log-spectral distortion; MAP, maximum a posteriori; MFMAPEA, multiframe MAP estimations of the amplitude; MFWF, multiframe Wiener filter; SDW-IFWF, speech-distortion weighted interframe Wiener filter; SegSNR, segmental SNR; SNR, signal-to-noise ratio; STOI, short-time objective intelligibility.

estimators) using the interface correlation concept. On the other side, compared with the multiframe beamforming-based algorithms, such as MFWF or SDW-IFWF, at least when the exact VAD is available, the proposed multiframe MAP-based algorithms achieve less improvement, especially for low SNRs.

Figure 2 presents the performance results in the presence of different types of additive input noises, including 'stationary pink', 'non-stationary babble', 'low-frequency Factory', and 'high-frequency Printer' noises. In addition, to provide a more realistic scenario, we used real recorded signals (from [28]), where the clean speech signals were selected from the public audiobook database, LibriVox, and the noise signals were generated using the formal mean opinion score (MOS)-like testing. The links to realistic scenarios and real recorded sample noisy and enhanced files are available at https://pws.yazd.ac.ir/sprl/Ranjibaryan-IET-SampleWaves.

4.2 Only noisy signals available

In this case, we possess only the noisy signals. To apply JMAP estimation, \( \sigma^2_e(m) \) was computed using the improved minima controlled recursive averaging method [30]. However, to update the noise correlation matrix, we use the recursive approach that is mentioned in [24], where only noisy signals are available. In multiframe case, the noise correlation matrix is given by

\[
\Phi_v = \lambda_n \Phi_{v,n-1} + (1 - \lambda_n) E\{v_m y_m^H | y_m \}
\]

where

\[
E\{v_m y_m^H | y_m \} = \frac{1}{1 + \delta_m} y_m y_m^H + \frac{\delta_m}{1 + \delta_m} \Phi_v
\]
\( \lambda_n \) denotes the forgetting factor for noise, and \( \hat{\lambda}_n \) represents the a priori SNR as defined in Equation (49).

It should be noted that noise statistics do not usually change drastically in successive frames. Consequently, we can utilize \( \Phi_{x_{n-1}} \) instead of \( \Phi_{x_n} \) in Equation (51). We assume that the first 10 frames consist of noise only. Therefore, we use these frames to obtain initial estimates of the correlation matrix of the noise signal. In our extensive experiments, we found that the best performance is achieved by choosing \( \lambda_y = \lambda_x = 0.9 \).

In addition, in the following, taking the SPP into account, we consider a mixed scenario. According to the low-pass nature of speech signal, we consider the interframe correlation for the cases where only speech is present \( (SPP(m) = 1) \). On the other side, in the case of silent frames \( (SPP(m) = 0) \), we ignore the multiframe correlation and apply the JMAP estimator. The SPP-based estimator is given by

\[
\hat{X}_m = SPP(m) \hat{X}_{MF}(m) + (1 - SPP(m)) \hat{X}_{JMAP}(m). \tag{52}
\]

where \( SPP(m) \) denotes the SPP at the current time frame, \( \hat{X}_{MF}(m) \) denotes the estimated clean speech signal provided by multiframe methods (multiframe MAP estimators and Wiener filters), and \( \hat{X}_{JMAP}(m) \) represents the estimated clean speech signal using JMAP estimator.

To compute the SPP(m), we used the single-microphone SPP estimator proposed in [11] that utilizes both interframe and interband correlations.

Using Equation (52), Table 2 depicts the performance of the considered SPP-based algorithms in the presence of additive white Gaussian noise at input full-band SNRs ranging from -10 to 15 dB. We have highlighted the best performance for each SNR in bold.

In terms of PESQ, the JMFMAP-\( \beta \) estimator delivers the best performance. The next ranks are for the MFMAPEA-Amp and MFWF, which their performances are very close.
It was shown that the performance of the multiframe algorithms (either MFMAP or MFWFs) is highly sensitive to the accurate estimation of the noise correlation matrix. Indeed, the highly time-varying IFC leads to an inaccurate correlation matrix. This leads to the conclusion that further investigation to improve the robustness of multiframe estimators against noise correlation matrix is needed. It also observed that when an accurate estimation of $\beta$-parameter is available, the multiframe MAP estimator has the potential to improve the PESQ values.

Regarding STOI improvement, it is seen that the comparative performance of the estimators can be sorted as JMFMAP-\(\beta\), MFMAPEA-Sig and JMFMAP (very close), JMAP, MFWF, and MFMAPEA-Amp (very close), respectively, in low SNRs. In high SNRs, JMAP delivers the best improvement for white Gaussian noise.

In terms of SegSNR, among the multiframe MAP estimators, the JMFMAP-\(\beta\) outperforms the rest for all input SNRs. Next is the MFMAPEA-Amp. The performance of MFWF and SDW-IFWF lie in between JMFMAP and MFMAPEA-Sig in mid and high SNRs. It is also seen that the JMAP provides the greatest improvement in high SNRs.

Concerning the LSD values, the results for all considered estimators are very close, indicating the MFWF delivers slightly better results than the others in low SNRs and JMAP in high SNRs.

5 | CONCLUSION

In this work, the concept of statistical speech properties has been included in the context of interframe correlation in time-frequency analysis. Assuming Gaussian statistical modelling of the speech and noise signals, several multiframe MAP estimators in a single-microphone case have been introduced. In each TFU, a vector consisting of the observations at the current frame and a finite number of past frames was considered to derive estimators. Consequently, closed-form solutions for both phase and amplitude of a clean speech signal were provided, relying only on second-order statistical properties.

In this contribution, the effect of different numbers of frames was investigated. Simulation results confirmed that it is possible to improve the performance of the existing single-frame Bayesian methods using the interframe correlation concept. In addition, the performances of the proposed multiframe MAP estimators with those of JMAP, MFWF, and SDW-IFWF for various kinds of noises were compared. It was observed that similar to the multiframe beam-forming methods, our proposed multiframe MAP estimators deliver significant improvement if the noise correlation matrix is accurately estimated. In addition, using real recorded signals, the practical efficiency of considered estimators was demonstrated in realistic scenarios.

### Table 2

| Methods/\Input SNR (dB) | PESQ Improvement | \(\beta\) | 0 | 5 | 10 | 15 | Methods/\Input SNR (dB) | PESQ Improvement | \(\beta\) | 0 | 5 | 10 | 15 |
|------------------------|------------------|---------|---|---|---|---|------------------------|------------------|---------|---|---|---|---|
| JMAP                   | 0.40             | 0.52   | 0.65 | 0.67 | 0.64 | 0.56 | JMAP                   | -0.01           | 0.000  | 0.01 | 0.011 | 0.001 | -0.00 |
| JMFMAP                 | 0.43             | 0.57   | 0.67 | 0.67 | 0.62 | 0.51 | JMFMAP                 | -0.01           | 0.004  | 0.02 | 0.010 | -0.008 | -0.02 |
| JMFMAP-\(\beta\)       | 0.56             | 0.71   | 0.83 | 0.82 | 0.77 | 0.68 | JMFMAP-\(\beta\)       | 0.02            | 0.037  | 0.05 | 0.031 | 0.005 | -0.01 |
| MFMAPEA-Sig            | 0.43             | 0.57   | 0.67 | 0.69 | 0.66 | 0.56 | MFMAPEA-Sig            | -0.00           | 0.004  | 0.02 | 0.009 | -0.006 | -0.02 |
| MFMAPEA-Amp            | 0.43             | 0.59   | 0.71 | 0.70 | 0.66 | 0.58 | MFMAPEA-Amp            | -0.01           | -0.003 | 0.01 | 0.007 | -0.007 | -0.01 |
| MFWF                   | 0.43             | 0.58   | 0.70 | 0.69 | 0.66 | 0.58 | MFWF                   | -0.01           | -0.002 | 0.01 | 0.006 | -0.009 | -0.02 |
| SDW-IFWF               | 0.43             | 0.57   | 0.70 | 0.70 | 0.67 | 0.55 | SDW-IFWF               | -0.02           | -0.005 | 0.01 | 0.009 | -0.003 | -0.01 |

SegSNR improvement: LSD

JMAP: 13.45 12.00 10.29 8.50 6.71 5.09 10.79 8.40 6.22 4.61 3.41 2.44
JMFMAP: 13.14 11.50 9.30 6.32 2.70 -1.32 11.16 8.64 6.44 4.80 3.64 2.81
JMFMAP-\(\beta\): 14.60 13.09 11.24 8.93 6.13 2.91 10.3 7.92 5.89 4.43 3.38 2.60
MFMAPEA-Sig: 13.17 11.77 10.03 7.85 5.17 2.06 11.15 8.60 6.38 4.76 3.61 2.78
MFMAPEA-Amp: 13.98 12.37 10.43 8.11 5.37 2.28 10.04 7.88 6.00 4.60 3.54 2.71
MFWF: 14.41 12.45 10.01 7.04 3.63 0.00 9.67 7.63 5.83 4.51 3.54 2.30
SDW-IFWF: 14.16 12.31 9.98 7.14 3.82 0.26 9.76 7.72 5.90 4.56 3.55 2.76

Note: Bolded items show the best performance in each SNR.

Abbreviations: JMAP, joint MAP; JMFMAP, joint multiframe MAP; LSD, log-spectral distortion; MAP, maximum a posteriori; MFMAPEA, multiframe MAP estimations of the amplitude; MFWF, multiframe Wiener filter; SDW-IFWF, speech-distortion weighted interframe Wiener filter; SegSNR, segmental SNR; SNR, signal-to-noise ratio; STOI, short-time objective intelligibility.
This work can be considered a starting point for combining both statistical speech properties and IFC concepts. One of the significant advantages of this work is extending and applying different distributions for clean speech and noise signals. It would also be interesting to investigate hybrid approaches that employ deep learning techniques to obtain noise statistics while assuming Gaussian or super-Gaussian distributions for modeling speech signals.

ACKNOWLEDGEMENTS

We would like to express our appreciation to Iran National Science Foundation (INSF) for supporting this work under Grant number 96000455.

ORCID

Hamid Reza Abutalebi https://orcid.org/0000-0002-1408-5637

REFERENCES

1. Benesty, J., Chen, J., Huang, Y.: Microphone Array Signal Processing. Springer-Verlag Berlin Heidelberg, Berlin (2008)
2. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32(6), 1109–1121 (1984)
3. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33(2), 443–445 (1985)
4. Wolfe PJ, Godsell SJ: Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. EURASIP J. Adv. Signal Process. 2003(10), 910167 (2003). http://dx.doi.org/10.1155/s11108657030340111
5. You, C.H., Koh, S.N., Rahardja, S.: β‐order MMSE spectral amplitude estimation for speech enhancement. IEEE Trans. Speech Audio Process. 13(4), 475–486 (2005)
6. Lotter, T., Vary, P.: Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. EURASIP J. Adv. Signal Process. 2005(7), 1110–1126 (2005)
7. Hendriks, R.C., Martin, R.: MAP estimators for speech enhancement under normal and Rayleigh inverse Gaussian distributions. IEEE Trans. Audio Speech Lang. Process. 15(3), 918–927 (2007)
8. Lotter, T., Benen, C., Vary, P.: Multi channel direction independent speech enhancement using spectral amplitude estimation. EURASIP J. Adv. Signal Process. 2003(11), 1147–1156 (2003)
9. Trawicki, M.B., Johnson, M.T.: Distributed multichannel speech enhancement with minimum mean-square error short-time spectral amplitude, log-spectral amplitude, and spectral phase estimation. Signal Process. 92(2), 345–356 (2012)
10. Huang, Y.A., Benesty, J.: A multi-frame approach to the frequency-domain single-channel noise reduction problem. IEEE Trans. Audio Speech Lang. Process. 20(4), 1256–1269 (2012)
11. Momeni, H., Habets, E.A.P., Abutalebi, H.R.: Single-channel speech presence probability estimation using inter-frame and inter-band correlations. In: Proceedings on International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2903–2907. Florence (2014)
12. Schasse, A., Martin, R.: Estimation of subband speech correlations for noise reduction via MVDR processing. IEEE/ACM Trans. Audio Speech Lang. Process. 22(9), 1355–1365 (2014)
13. Momeni, H., Abutalebi, H.R., Habets, E.A.P.: Conditional MMSE-based single-channel speech enhancement using inter-frame and inter-band correlations. In: Proceedings on International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5215–5219. Shanghai (2016)
14. Ranjbaryan, R., Abutalebi, H.R.: Multi-channel estimation of power spectral density matrix using inter-frame and inter-band information. In: Proceedings on European Signal Processing Conference (EUSIPCO), pp. 1684–1688. Kos (2017)
15. Fischer, D., Docolo, S.: Sensitivity analysis of the multi-frame MVDR filter for single-microphone speech enhancement. In: Proceedings on European Signal Processing Conference (EUSIPCO), pp. 603–607. Kos (2017)
16. Andersen, K.T., Moonen, M.: Robust speech-distortion weighted inter-frame Wiener filters for single-channel noise reduction. IEEE/ACM Trans. Audio Speech Lang. Process. 26(1), 97–107 (2018)
17. Kazama, M., et al.: On the significance of phase in the short term Fourier spectrum for speech intelligibility. J. Acoust. Soc. Am. 127(3), 1432–1439 (2010)
18. Gerkmann, T., Krawczyk Becker, M., Roux, J.L.: Phase processing for single-channel speech enhancement. IEEE Signal Process. Mag. 32(2), 55–66 (2015)
19. Gerkmann, T.: MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase. In: Proceedings on International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4478–4482. Florence (2014)
20. Oo, N., Gan, W.: On Harmonic Addition Theorem. Int. J. Comput. Commun. Eng. 1(3), 200–202 (2012). http://dx.doi.org/10.7763/ijccee.2012.v1.52
21. Roy, R., Kailath, T.: ESPRIT-estimation of signal parameters via rotational invariance techniques. IEEE Trans. Acoust., Speech, Signal Process. 37(7), 984–995 (1989)
22. Loizou, P.: Speech Enhancement: Theory and Practice. CRC press. Florida (2007)
23. Tual, C.H., et al.: A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: Proceedings on International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4214–4217. Dallas (2010)
24. Gerkmann, T., Hendriks, R.C.: Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans Audio Speech Lang Process. 20(4), 1383–1393 (2012)
25. Abramson, A., Cohen, I.: Simultaneous detection and estimation approach for speech enhancement. IEEE Trans Audio Speech Lang Process. 15(8), 2348–2359 (2007)
26. Garofolo, J.S.: Getting Started With the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database. National Institute of Standards and Technology (NIST), Gaithersburg (1993). https://doi.org/10.6028/nist.ir.4930
27. Mowlaee, P., Gaich, A.: On speech intelligibility estimation of phase-aware single-channel speech enhancement. In: Proceedings of the International Speech Communication Association IS 10 (2015)
28. Reddy, C.K., et al.: The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results, arXiv preprint arXiv:200513981. Shanghai (2020)
29. Goumerek, J.F., et al.: Audio set: an ontology and human-labelled dataset for audio events. In: Proceedings on International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 776–780. New Orleans (2017)
30. Cohen, I.: Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. 11(5), 466–475 (2003)

How to cite this article: Ranjbaryan R, Abutalebi HR. Multiframe maximum a posteriori estimators for single-microphone speech enhancement. IET Signal Process. 2021;1–15. https://doi.org/10.1049/sil2.12045
APPENDIX: ESPRIT algorithm
Consider two subarrays, each including \( J \) sensors. Assume that \( d \) uncorrelated signals with directions \( \theta_1, \ldots, \theta_d \) arrived at the subarrays. The received signals can be written as follows:

\[
\begin{align*}
\mathbf{u}_1 &= \mathbf{G}(\theta)\mathbf{s} + \mathbf{v}_1, \\
\mathbf{u}_2 &= \mathbf{G}(\theta)\mathbf{F}(\theta)\mathbf{s} + \mathbf{v}_2,
\end{align*}
\]  

(53)

Note that, \( \mathbf{u}_1 \) and \( \mathbf{u}_2 \in \mathbb{C}^{J \times 1} \), the steering matrix \( \mathbf{G}(\theta) \in \mathbb{C}^{J \times d} \), the signal vector \( \mathbf{s} \in \mathbb{C}^{d \times 1} \), and the additive noises \( \mathbf{v}_1 \) and \( \mathbf{v}_2 \in \mathbb{C}^{J \times 1} \). The matrix \( \mathbf{F}(\theta) \in \mathbb{C}^{d \times d} \) indicates a diagonal matrix representing phase delays between subarrays, which is given by \( \mathbf{F}(\theta) = \text{diag}(e^{\delta_1}, \ldots, e^{\delta_d}) \), where \( \delta_j \) is a function of DOA, frequency and so on. The aim of ESPRIT algorithm is to compute the matrix \( \mathbf{F}(\theta) \) and consequently the corresponding direction of arrivals, \( \theta_1, \ldots, \theta_d \) [21].

In [21], it was shown that when \( d \leq J \), the ESPRIT algorithm is able to estimate matrix \( \mathbf{F}(\theta) \). To this purpose, authors defined a concatenation vector as

\[
\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{G}(\theta) \\ \mathbf{G}(\theta)\mathbf{F}(\theta) \end{bmatrix} \mathbf{s} + \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = \mathbf{H}(\theta)\mathbf{s} + \mathbf{v},
\]  

(54)

where \( \mathbf{H}(\theta) = \begin{bmatrix} \mathbf{G}(\theta) \\ \mathbf{G}(\theta)\mathbf{F}(\theta) \end{bmatrix} \) and \( \mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} \). The correlation matrix of \( \mathbf{u} \) is given by

\[
\mathbf{\Phi}_u = \mathbf{H}(\theta)\mathbf{\Phi}_s\mathbf{H}^H(\theta) + \mathbf{\Phi}_v
\]  

(55)

In [21], it was shown that when \( d \leq J \) uncorrelated signals arrived at subarrays, the \( d \) generalized eigenvectors (GEVs) of \( \mathbf{\Phi}_u, \mathbf{\Phi}_v \), corresponding to the \( d \) largest generalized eigenvalues, can be used to span the signal subspace. In other word, the signal subspace,

\[
\mathbf{E}_x = \mathbf{\Phi}_v\mathbf{e}_1 \cdots \mathbf{e}_d
\]  

(56)

is obtained by the corresponding GEV’s, \( \mathbf{e}_i, i = 1, \ldots, d \), where \( \mathcal{R}\{\mathbf{E}_x\} = \mathcal{R}\{\mathbf{H}(\theta)\} \) (\( \mathcal{R} \) denotes range).

Because \( \mathbf{E}_x \) and \( \mathbf{H}(\theta) \) span the same subspace \( \mathcal{R}\{\mathbf{H}(\theta)\} = \mathcal{R}\{\mathbf{H}(\theta)\} \), a unique and non-singular matrix \( \mathbf{T} \) exist such that [21].

\[
\mathbf{E}_x = \mathbf{H}(\theta)\mathbf{T}.
\]  

(57)

Partitioning \( \mathbf{E}_x \) into \( \mathbf{E}_x \in \mathbb{C}^{J \times d} \) and \( \mathbf{E}_y \in \mathbb{C}^{J \times d} \) we have

\[
\begin{bmatrix} \mathbf{E}_x \\ \mathbf{E}_y \end{bmatrix} = \begin{bmatrix} \mathbf{G}(\theta)\mathbf{T} \\ \mathbf{G}(\theta)\mathbf{F}(\theta)\mathbf{T} \end{bmatrix}.
\]  

(58)

It is seen that \( \mathcal{R}\{\mathbf{E}_x\} = \mathcal{R}\{\mathbf{E}_y\} = \mathcal{R}\{\mathbf{G}(\theta)\} \), and consequently span a same column space. As a result, rank of \( \mathbf{E}_{xy} = \begin{bmatrix} \mathbf{E}_x & \mathbf{E}_y \end{bmatrix} \) is equal to \( d \). Because the rank of the \( J \times 2d \) dimensional matrix \( \mathbf{E}_{xy} \) is \( d \), there must exists a unique matrix \( \mathbf{W} \) spanning the null space as follows [21].

\[
0 = \begin{bmatrix} \mathbf{E}_x\mathbf{E}_y \end{bmatrix} \mathbf{W} = \mathbf{E}_x\mathbf{W}_x + \mathbf{E}_y\mathbf{W}_y = \mathbf{G}(\theta)\mathbf{T}\mathbf{W}_x + \mathbf{G}(\theta)\mathbf{F}(\theta)\mathbf{T}\mathbf{W}_y
\]  

(59)

defining \( \mathbf{\Psi} = -\mathbf{W}_x\mathbf{W}_y^{-1} \)

\[
\mathbf{G}(\theta)\mathbf{T}\mathbf{\Psi} = \mathbf{G}(\beta)\mathbf{F}(\theta)\mathbf{T} \Rightarrow \mathbf{G}(\theta)\mathbf{T}\mathbf{\Psi}\mathbf{T}^{-1} = \mathbf{G}(\theta)\mathbf{F}(\theta)
\]  

(60)
implies that

\[
\mathbf{T}\mathbf{\Psi}\mathbf{T}^{-1} = \mathbf{F}(\theta)
\]  

(61)

As a consequence, the eigenvalues of \( \mathbf{\Psi} \) represent the diagonal elements of \( \mathbf{F}(\beta) \), and the eigenvectors of \( \mathbf{\Psi} \) represent the column of matrix \( \mathbf{T} \).

Briefly, the Total least square ESPRIT algorithm operates as follows [21]:

1. Define the concatenation vector and its correlation matrix as Equations (54) and (55), respectively.
2. Compute the generalized eigen-decomposition of \( (\mathbf{\Phi}_u, \mathbf{\Phi}_v) \) and consequently the \( d \) GEVs corresponding to the \( d \) largest GVs.
3. Compute the signal subspace using Equation (56) and decompose it to obtain \( \mathbf{E}_x \) and \( \mathbf{E}_y \) as \( \mathbf{E}_x = \begin{bmatrix} \mathbf{E}_x^1 \\ \mathbf{E}_x^2 \end{bmatrix} \).
4. Compute the matrix \( \mathbf{E} \) using the eigen-decomposition as follows

\[
\begin{bmatrix} \mathbf{E}_x^H \\ \mathbf{E}_y^H \end{bmatrix} \begin{bmatrix} \mathbf{E}_x^1 \\ \mathbf{E}_x^2 \end{bmatrix} = \mathbf{E}\mathbf{E}_x
\]  

(62)

5. Partition matrix \( \mathbf{E} \) into \( d \times d \) submatrices

\[
\begin{bmatrix} \mathbf{E}_{11} & \mathbf{E}_{12} \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{bmatrix}
\]  

(63)

6. Calculate \( \mathbf{\Psi} = -\mathbf{E}_{12}\mathbf{E}_{22}^{-1} \) and its eigenvalues.

The eigenvalues of \( \mathbf{\Psi} \) represent the diagonal elements of \( \mathbf{F}(\beta) \).

It should be noted that in our problem, the noisy observed signal at each TFU is considered as one subarray. Indeed

\[
\begin{align*}
\mathbf{u}_1 &\leftarrow \mathbf{Y}_m \\
\mathbf{u}_2 &\leftarrow \mathbf{Y}_{m-2}
\end{align*}
\]  

(64)

also, \( J = d = 1 \). As mentioned before, when \( d = J \), the ESPRIT algorithm is still able to find the phase difference. As it is seen \( \mathbf{F}(\beta) \leftarrow C_2m e^{-j\beta_{2m}} \).