A New Observation Model in the Logarithmic Mel Power Spectral Domain for the Automatic Recognition of Noisy Reverberant Speech

Alexander Krüger, Member, IEEE, Volker Leutnant, Student Member, IEEE and Reinhold Haeb-Umbach, Senior Member, IEEE

Abstract—In this contribution we present a theoretical and experimental investigation into the effects of reverberation and noise on features in the logarithmic mel power spectral domain, an intermediate stage in the computation of the mel frequency cepstral coefficients, prevalent in automatic speech recognition (ASR). Gaining insight into the complex interaction between clean speech, noise, and noisy reverberant speech features is essential for any ASR system to be robust against noise and reverberation present in distant microphone input signals. The findings are gathered in a probabilistic formulation of an observation model which may be used in model-based feature compensation schemes. The proposed observation model extends previous models in three major directions: First, the contribution of additive background noise to the observation error is explicitly taken into account. Second, an energy compensation constant is introduced which ensures an unbiased estimate of the reverberant speech features, and, third, a recursive variant of the observation model is developed resulting in reduced computational complexity when used in model-based feature compensation. The experimental section is used to evaluate the accuracy of the model and to describe how its parameters can be determined from test data.

Index Terms—robust automatic speech recognition, model-based feature compensation, observation model for reverberant and noisy speech, recursive observation model

I. INTRODUCTION

HANDS-FREE systems have the potential to improve the acceptance of automatic speech recognition in many application areas. Allowing the user to move freely without the need of wearing a head-set or holding a microphone, they contribute to increased convenience as well as safety.

However, the increased distance of the speaker to the microphone compared to the use of head-sets leads to an acoustical signal degradation due to reverberation and background noise, and thus, as a consequence, to an increased word error rate of an ASR system. Even techniques known to be effective against additive noise, e.g., those used in the ETSI Advanced Front-End [1], are rather ineffective towards reverberation [2].

Key to the success of any approach to reverberant speech recognition is an understanding of how reverberation affects speech. This may be expressed by a so-called observation model, which relates the observed signal to the underlying clean non-reverberant signal. The observation model is the core component of model-based feature compensation, a paradigm to environmentally robust ASR, which has lead to a variety of very effective noise compensation algorithms.

The goal of this paper is to derive a powerful and yet computationally tractable observation model for noisy reverberant speech, eventually laying the ground to extend model-based feature compensation to the robust recognition of speech which is corrupted by both noise and reverberation.

A system theoretic model for the effect of reverberation is the convolution of the source signal $\varphi(l)$ with the acoustic impulse response (AIR) $h_l(p')$ from the source to the sensor. In the additional presence of background noise $\mu(l)$, which is usually present if speech is captured by distant microphones, the microphone signal $y(l)$ is given by

$$y(l) = \sum_{p'=0}^{L_h-1} h_l(p') \varphi(l-p') + \mu(l). \quad (1)$$

The AIR of assumed length $L_h$ is in general time-variant. Here, the subscript $l$ indicates the time variance whereas $p' \in \mathbb{N}_0$ is the lag index. The time variance of the AIR is due to changes within the source-sensor enclosure, e.g., due to movements of the speaker, movements in the environment or changes in temperature.

Depending on where the effect of reverberation is addressed, the observation model has to be formulated as a relationship among the signals, the features or the statistical models corresponding to the clean data, the noise and the corrupted observations.

The model (1) is at the outset of many speech enhancement algorithms which aim at dereverberating the signal by inverse filtering [3], [4], blind deconvolution [5]–[7], or by multi-step linear prediction [8]. Most of these techniques assume availability of multi-channel data. As this may be an unrealistic assumption for many applications, we will in the following consider approaches that call for only a single microphone.

As the estimation of the AIR is a complicated task, simplified observation models have been proposed which do not require the estimation of the full AIR but rather some of its characteristics. They represent the effect of reverberation in the spectral or log-spectral domain rather than in the time domain. The proposed simplified observation models can be broadly categorized into three groups:

i) Models that assume a linear, affine or additive relationship between clean and reverberant features in the logarithmic mel power spectral or cepstral domain,
neglecting any temporal correlation introduced by reverberation [9]–[13].

ii) Models that describe the reverberation as an additive distortion in the power spectral domain [14]–[18].

iii) Models that describe reverberation by a convolution in the power spectral domain [19]–[22].

Clearly, the models falling into the third category are closest to the physical model of Eq. (1).

The observation model to be presented in this paper falls also in this category. Although not discussed here, the model is meant to be employed for Bayesian feature enhancement whereby the posterior PDF of the non-reverberant and noise-free logarithmic mel power spectral feature vector is estimated from the noisy reverberant input and then forwarded to an automatic speech recognition back end. By enhancing the features rather than the signal one can take advantage of both a decimation in time, since the frame rate at which feature vectors are computed is much lower than the sampling rate, and in frequency, due to the mel filter bank applied to the power spectrum. This simplifies the estimation of a frequency domain representation of the AIR. Further, it is generally considered advantageous if the representation to be enhanced, here the features rather than the acoustic signal, is close to what is actually processed in the recognizer. The enhancement can then be tailored to the specifics of the recognizer rather than to those of a human listener.

For the absence of additive background noise an approximation for the relation between the mel power spectral coefficients (MPSCs) of the reverberant speech signal and that of the clean speech signal was presented in [21]. Later we will show that due to the absence of a power compensation factor, the model, however, suffers from a systematic underestimation of the power spectrum of reverberant speech. A convolutive observation model was also employed in the static acoustic model adaptation technique presented in [20]. In [19] a recursive observation model was employed for dynamically adapting an acoustic model to reverberation. It can be viewed as a simplified version of the one presented here.

With the aforementioned models the feature vectors of the observable reverberant and noisy speech signal are expressed by means of the feature vectors of the noise and the noise-free and non-reverberant speech in a purely deterministic way. However, due to loss of information inherent in the process of the feature computation, the formulation of an exact deterministic relationship in the feature domain is impossible in general. Instead of neglecting the remaining error between the model and the true observation, an improved modeling may be achieved by describing this error (and thus the observation) in a probabilistic way. Such a probabilistic model has been presented in [18], where reverberation was described as additive in the mel power spectral domain (category ii) above). In [22] we developed such a model for the logarithmic mel power spectral domain, however assuming absence of the additive noise term \( \pi(l) \). Later, the model was extended to the presence of additive noise [2]. However, that extension only considered the contribution of the noise term to the deterministic part of the observation model, while its effect on the observation error was neglected. Here, we extend this model and present a more refined treatment of the observation error, including the contribution of additive noise.

It is well known that modeling the effect of additive noise as additive in the power spectral domain is only an approximation, which breaks down at signal-to-noise ratios (SNRs) close to 0 dB. Then, the cross term in the computation of the power spectrum of a signal consisting of a superposition of speech and noise can no longer be neglected [23]. This term is significantly more complicated in the presence of reverberation. The reason is quite simple: reverberation results from filtering the speech signal by the AIR and is thus highly non-stationary. Consequently, the cross term becomes highly non-stationary and therefore difficult to model. This paper presents, to our knowledge, for the first time an observation model for reverberant speech that includes the impact of the cross term. Our aim is to strike a balance between model accuracy and tractability. While the description of its use for feature enhancement and subsequent speech recognition is beyond the scope of this paper, it is worthwhile mentioning that the new model led to significantly improved ASR results at low SNRs [24].

The organization of the paper is as follows. First, in Sec. II we present the assumed model for the corrupted speech signal and review the individual steps in the computation of the MFCCs to introduce the notation, which will be essential for the derivation of the observation models in Sec. III. In Sec. IV we access the properties of the derived observation model experimentally. In particular we evaluate the statistical properties of the observation error and the influence of the power compensation constant. We also discuss how to determine the parameters of the observation model. The paper is finally concluded by Sec. V.

II. SIGNAL MODEL AND REVIEW OF FEATURE EXTRACTION

Neglecting the time variance of the AIR, Eq. (1) can be written as

\[
\overline{y}(l) = \sum_{p' = 0}^{L_h - 1} h(p') \overline{x}(l - p') + \overline{n}(l) =: \overline{x}(l) + \overline{n}(l),
\]

where \( \overline{n}(l) \) denotes the reverberant speech signal.

In the following, we recapitulate the extraction of the MFCCs from the discrete-time microphone signal \( \overline{y}(l) \) according to a slightly modified version of the ETSI Standard Front-End [25], which is illustrated in Fig. 1. The modifications are the replacement of the magnitude spectrum by the power spectrum and the replacement of the logarithmic frame energy by the zeroth cepstral component.

For the feature extraction, the microphone signal \( \overline{y}(l) \) is first passed to an offset compensation and a pre-emphasis stage. The resulting discrete-time signal \( y(l) \) is framed and windowed by an analysis window \( w_A(l) \) of length \( L_w \), which fulfills

\[
w_A(l) = 0 \quad \text{for} \quad l < 0 \quad \land \quad l > L_w
\]

and which is shifted by the frame shift \( B \) between consecutive frames. The resulting windowed frames

\[
y_w(m, l') := w_A(l') y(l' + mB),
\]
where \( m \) denotes the frame index and \( l' \) denotes the discrete time index within each frame, are subsequently transformed to the frequency domain by applying a discrete Fourier transform (DFT):
\[
Y(m,k) := \sum_{l'=0}^{L_w-1} y_{w_A}(m,l') \cdot e^{-j \frac{2 \pi}{K_S} kl'}.
\] (5)

Here \( k \in \{0, \ldots, K - 1\} \) is the frequency bin index, \( K \) is the number of frequency bins, and \( j \) denotes the imaginary unit.

The mel power spectral coefficients (MPSCs) are computed by integrating the triangularly weighted power spectrum within each of a total of \( Q \) overlapping mel frequency bands, indexed by \( q \in \{0, \ldots, Q - 1\} \) according to
\[
Y_{m,q} := \sum_{k=K_{q}^{(0)}}^{K_{q}^{(t)}} |Y(m,k)|^2 \Lambda_q(k).
\] (6)

The center frequencies for the corresponding triangular weighting functions \( \Lambda_q(k) \) are equally spaced on the mel scale. The width of each mel band is given by the difference of the corresponding upper and lower bounds, \( K_{q}^{(0)} \) and \( K_{q}^{(t)} \), respectively.

The next step is a compression of the mel spectrum by applying the natural logarithm resulting in the logarithmic MPSCs (LMPSCs)
\[
y_{m,q} := \ln \{Y_{m,q}\}.
\] (7)

Finally, the MFCCs are obtained by applying a discrete cosine transform (DCT) to the LMPSCs. For ease of notation in sections following, we additionally introduce the LMPSC vector
\[
y_m := (y_{m,0}, \ldots, y_{m,q-1})^T,
\] (8)

which is formed by collecting the LMPSCs of all mel bands for a fixed frame index \( m \). All vectors introduced in the following sections will be composed of their individual mel components in an analogous way.

### III. Observation models

In this section we are going to derive a probabilistic relationship between the observed LMPSC vector \( y_m \), the underlying sequence of clean, non-reverberant speech LMPSC vectors \( x_m, x_{m-1}, \ldots \) and the noise LMPSC vector \( n_m \). This may be employed, e.g., to obtain the observation PDF, i.e., the conditional probability density of \( y_m \), given the clean speech LMPSC vectors and the noise LMPSC vector, which is required for model-based feature compensation.

We start with a brief review of the observation model in the absence of noise as derived in [22] and will develop it further to also account for additional background noise.

#### A. Observation model for absence of background noise

In the absence of background noise, the discrete-time microphone signal \( \pi(l) \) is given by the convolution of the clean speech signal \( \pi(l) \) with the AIR \( h(l) \) as given in (2). According to [22], the corresponding STDFT \( S(m,k) \) may be expressed as
\[
S(m,k) = \sum_{k'=0}^{K-1} \sum_{m'=0}^{L_H} X(m-m',k')h_{k,k'}(m')
\] (9)

with
\[
h_{k,k'}(m') := \sum_{p'=0}^{L_h-1} h(p')\phi_{k,k'}(m'B-p')
\] (10)

and
\[
\phi_{k,k'}(l) := e^{j \frac{2 \pi}{K} k l} \sum_{l'=0}^{L_w-1} w_A(l')w_S(l'+l)e^{-j \frac{2 \pi}{K} (k-k')l'}.
\] (11)

The terms \( h_{k,k'}(m') \) will in the following be referred to as cross-band filters for \( k \neq k' \) and as band-to-band filters for \( k = k' \), as in [26]. The lengths \( L_{H,L} \) and \( L_H \) in (9) are defined by
\[
L_{H,L} := \left[ \frac{L_w - 1}{B} \right], \quad L_H := \left[ \frac{L_h + L_w - 2}{B} \right].
\] (12)

Further, \( w_S(l') \) denotes a synthesis window, which is bi-orthogonal to \( w_A(l') \) [27] and has the same support as \( w_A(l') \). For the power of \( S(m,k) \) we now write
\[
|S(m,k)|^2 = C_P \sum_{m'=0}^{L_H} |X(m-m',k)|^2 |h_{k,k'}(m')|^2 + E(S)(m,k).
\] (13)

The term \( E(S)(m,k) \) thereby captures the error of approximating the square of the sum given in (9) by the sum of the squares while also ensuring a causal relationship by dropping all negative frame indices. The constant \( C_P \) will be determined such that the error term \( E(S)(m,k) \) is zero-mean. Introducing the expectation operator \( E[\cdot] \), this, however, is equivalent to
\[
E \left[ |\hat{S}(m,k)|^2 \right] = E \left[ C_P \sum_{m'=0}^{L_H} |\hat{X}(m-m',k)|^2 |\hat{h}_{k,k'}(m')|^2 \right].
\] (14)

Note that in (14) and in the following, we use the breve mark (˘) to distinguish a random variable from its realization. Due to the role of the constant \( C_P \) in (14), it is in the following
referred to as the power compensation constant. Besides being additive in the targeted LMPSC domain, choosing a multiplicative constant $C_P$ rather than an additive term to compensate for the bias introduced by the approximation is advantageous, since the desired compensation is made independent of the power of the clean speech signal and that of the AIR. This can best be seen by looking at the computation of $C_P$, addressed in Sec. III-D. By further introducing the mean of $|h_{k,k}(m')|^2$ over the $q$th mel band, i.e.,

$$\tilde{H}_{m',q} := \frac{1}{K_q} \sum_{k=K_q}^{K_u} |h_{k,k}(m')|^2,$$

(15)

the MPSCs $S_{m,q}$ of the reverberant speech signal can be written as

$$S_{m,q} = C_P \sum_{m'=0}^{L_H} \tilde{H}_{m',q} X_{m-m',q} + \mathcal{E}_{m,q}^{(S)}$$

$$=: \tilde{S}_{m,q} + \mathcal{E}_{m,q}^{(S)}.$$  

(16)

Hereby $X_{m,q}$ denote the MPSCs of the clean speech signal and $\mathcal{E}_{m,q}^{(S)}$ the error resulting from the approximation of $S_{m,q}$ by $\tilde{S}_{m,q}$. By introducing the logarithmic mel power spectral representation of the AIR

$$\tilde{h}_{m',q} := \ln \{ \tilde{H}_{m',q} \},$$

(18)

and the LMPSC of the clean speech signal

$$x_{m,q} := \ln \{ X_{m,q} \},$$

(19)

we are now able to express the MPSCs of the reverberant speech in terms of the underlying LMPSCs of the clean speech and the MPSCs of the AIR, i.e.,

$$s_{m,q} := \ln \{ S_{m,q} \} = \tilde{s}_{m,q} + v_{m,q}^{(s)},$$

(20)

where

$$\tilde{s}_{m,q} := \ln \{ \tilde{S}_{m,q} \} = \ln \left\{ C_P \sum_{m'=0}^{L_H} e^{x_{m-m',q}+\tilde{h}_{m',q}} \right\}. $$

(21)

Thereby, employing the definition of $\tilde{S}_{m,q}$ given in (17) instead of the equivalent formulation given in (21), the additive observation error in the LMPSC domain

$$v^{(s)}_{m,q} := s_{m,q} - \tilde{s}_{m,q} = \ln \{ S_{m,q} \} - \ln \{ \tilde{S}_{m,q} \}$$

(22)

$$=: \ln \left\{ \frac{S_{m,q}}{\sum_{m'=0}^{L_H} \tilde{H}_{m',q} X_{m-m',q}} \right\} - \ln \{ C_P \}$$

(23)

captures the errors from the approximation of $S_{m,q}$ by $\tilde{S}_{m,q}$ in the MPSC domain. Note that the choice of $C_P$ only affects the mean of the observation error $v^{(s)}_{m,q}$.

By introducing the observation mapping

$$f_s (x_{m-L_H:m}, \tilde{h}_{0:L_H}) := \ln \left\{ C_P \sum_{m'=0}^{L_H} e^{x_{m-m'}+\tilde{h}_{m'}} \right\},$$

(24)

where the mathematical operations are understood to be performed on the vectors component-wise, the relationship (20) may compactly be formulated by

$$s_{m} = f_s (x_{m-L_H:m}, \tilde{h}_{0:L_H}) + v^{(s)}_{m},$$

(25)

where $x_{m-L_H:m} := \{ x_{m-L_H}, ..., x_{y} \}$ and $\tilde{h}_{0:L_H} := \{ \tilde{h}, ..., \tilde{h}_{L_H} \}$ denote sequences of $L_H + 1$ Q-dimensional vectors. Recall that vectors in either sequence as well as $s_n$ and $v_n^{(s)}$ are composed of their individual mel components similar to $y_m$ in (8).

B. Observation model for presence of background noise

In the presence of background noise the LMPSC of the reverberant speech is given by [28]

$$y_{m,q} = \ln \left\{ S_{m,q} + N_{m,q} + 2\alpha_{m,q} \sqrt{S_{m,q}N_{m,q}} \right\}$$

$$= \ln \left\{ e^{y_{m,q} + v_{m,q}} + e^{n_{m,q}} + 2\alpha_{m,q} e^{\frac{y_{m,q} + v_{m,q} + n_{m,q}}{2}} \right\} ,$$

(26)

(27)

where $n_{m,q} := \ln \{ N_{m,q} \}$ is the LMPSC of the noise obtained from its MPSC $N_{m,q}$ and

$$\alpha_{m,q} := \frac{2 \sum_{k=K_q}^{K_u} \Lambda_k (k) \Re \{ S(m,k)N^*(m,k) \}}{\sqrt{S_{m,q}N_{m,q}}}$$

(28)

is the phase factor relating the LMPSC of the reverberant speech and that of the noise to the noisy reverberation observation [23]. Here, $\Re \{ \cdot \}$ and $(\cdot)^*$ denotes the real part operator and the conjugation, respectively. Further, $N(m,k)$ specifies the STDFT of the noise signal in frame $m$ and frequency bin $k$. To arrive at a formulation with an additive error term, similar to (20) in the noise free case, we write

$$y_{m,q} = \ln \left\{ e^{\tilde{s}_{m,q} + v_{m,q}} + v^{(s)}_{m,q} \right\} + v^{(y)}_{m,q}. $$

(29)

This time, the observation error $v^{(y)}_{m,q}$ consists of contributions from the error $v^{(s)}_{m,q}$ in the noise-free case and also from the (unknown) phase factor $\alpha_{m,q}$. This becomes evident by comparing (27) and (29). Solving for $v^{(y)}_{m,q}$ gives

$$v^{(y)}_{m,q} = \ln \left\{ 1 + \left( v^{(s)}_{m,q} - 1 \right) + \frac{1}{1 + e^{n_{m,q} - \tilde{s}_{m,q}}} \right\} + 2\alpha_{m,q} e^{\frac{v^{(s)}_{m,q} - n_{m,q}}{2}}.$$

(30)

The relative importance of the two contributions depends on the difference of the noise LMPSC $n_{m,q}$ to the speech-related LMPSC $s_{m,q}$ defined in (21). A detailed discussion on this issue will be deferred to the experimental section IV-D.

With the observation mapping

$$f_y (x_{m-L_H:m}, \tilde{h}_{0:L_H}, n_{m}) := \ln \left\{ C_P \sum_{m'=0}^{L_H} e^{x_{m-m'}+\tilde{h}_{m'}+n_{m}} \right\}$$

(31)

the relationship (29) may be formulated in vector notation as

$$y_{m} = f_y (x_{m-L_H:m}, \tilde{h}_{0:L_H}, n_{m}) + v^{(y)}_{m}. $$

(32)
C. Acoustic impulse response model

For the evaluation of (24) and (31) the logarithmic mel power spectral representation of the AIR $\hat{h}_{0,L_H}$ is required. In practice, however, this representation is usually unknown. To avoid a sensitive blind estimation of the AIR for a computation of $\hat{h}_{0,L_H}$, we have proposed in [22] to employ a stochastic AIR model, which has previously been introduced in [29]. According to this model, the AIR is regarded to be a realization of a stochastic process $\hat{h}(l)$ according to

$$\hat{h}(l) = \sigma_h \hat{v}(l) \chi_h(l) e^{-\frac{l}{\tau_h}}; \quad (33)$$

where $\hat{v}(l)$ is a zero-mean white GAUSSIAN stochastic process of unit power. The indicator function

$$\chi_h(l) := \begin{cases} 1 & \text{for } 0 \leq l \leq L_h - 1 \\ 0 & \text{else} \end{cases} \quad (34)$$

assures the AIR to be causal having finite length $L_h$. The term $e^{-\frac{l}{\tau_h}}$ causes an exponentially decaying envelope. The decay constant $\tau_h$ depends on the reverberation time $T_{60}$ through

$$\tau_h = \frac{T_{60}}{3 \ln(10) T_S}; \quad (35)$$

with $T_S$ denoting the sampling duration. The constant $\sigma_h$ may be used to control the AIR energy. The advantage of using this model is that it has only two parameters, i.e., $\tau_h$ and $\sigma_h$, which can be estimated more easily than the complete AIR. If working on artificially reverberated data, the estimation of $\sigma_h$ may even be superfluous, since the AIR is usually normalized to unit energy. The latter may also be assumed when working with true recordings of reverberant speech if the audio data are properly normalized, as described in [30].

Further, the presented stochastic model of the AIR also allows us to derive the recursive observation model presented in Sec. III-E. Based on the AIR model (33) a reasonable length $L_h$ may be determined in dependence on $\tau_h$ by

$$L_h = L_h(\tau_h) = \left\lfloor \frac{\tau_h}{2 \ln(\epsilon_h)} \right\rfloor; \quad (36)$$

which is obtained by minimizing the AIR length under the constraint that the relative energy of the neglected part of the AIR is smaller than $\epsilon_h$ [22].

It has been shown in [31] that the PDFs of the individual components $\hat{h}_{0,\alpha}$ of the logarithmic mel power spectral AIR representation $\hat{h}_{0,L_H}$ can be well modeled by GAUSSIANS.

Moreover, their respective means and variances have been found to only depend on the decay constant $\tau_h$, the energy term $\sigma_h$ and on parameters of the ETSI Standard Front-End.

As we have previously done in [22], we therefore propose to approximate the usually unknown logarithmic mel power spectral representation of the AIR $\hat{h}_{0,L_H}$ in both of the observation mappings (24) and (31) by its mean $\mu_{\hat{h}_{0,L_H}}$ under the AIR model (33). By doing so, the time-variance of the AIR due to, e.g., small movements of the speaker is absorbed through the observation error to a certain degree.

D. Computation of the power compensation constant

Having introduced the stochastic AIR model, we are able to compute the power compensation constant $C_P$ from condition (14) by using the AIR model and two additional assumptions. We assume the AIR and the clean speech signal to be mutually independent and the latter to be a realization of a zero-mean white GAUSSIAN stochastic process. Under these assumptions, the derived $C_P$ is given by

$$C_P = \frac{C_N}{C_D}; \quad (37)$$

where

$$C_N := K^2 \sum_{m',m''=-L_H}^{L_H} \sum_{l=0}^{L_w-1} w_A(l) w_S(l) \cdot \left[ \sum_{l'=-L_w+1}^{L_H} \chi_h(m'B-l') e^{-\frac{2(m''-m')}{\tau_h}} w_A^2(-l'+1) \right] \quad (38)$$

$$C_D := \sum_{l=0}^{L_H-1} \sum_{l'=0}^{L_w-1} w^2(-l') \chi_h(m'B-l') e^{-\frac{2(m''-m')}{\tau_h}} \quad (39)$$

and

$$w(l) := \sum_{l'=0}^{L_w-1} w_A(l') w_S(l'-l). \quad (40)$$

For a detailed derivation, the reader is referred to [32]. From (38) and (39), it can be seen that the constant $C_P$ only depends on the parameters employed for feature extraction and on the reverberation time. However, further analysis showed that there is only a weak dependence of $C_P$ on the latter [32].

E. Recursive observation model

In the presence of severe reverberation the length of the logarithmic mel spectral AIR representation, $L_H+1$, is large, such that many exponential terms occur in the observation mappings (24) and (31). In order to reduce the number of exponential terms, and hence to reduce the computational effort for their evaluation, we propose to employ a recursive observation model, which will be derived in the following.

The goal of such a recursive observation model is to express the observation at time instant $m$ in terms of the observation at a previous time instant $m-L_R$ and a reduced number of $L_R \ll L_H+1$ LMPSCs of the clean speech signal. For noisy reverberant observations, additional knowledge about the LMPSC of the noise at the current time instant $m$ and the previous time instant $m-L_Q$ will be used.

The basis for the derivation is the stochastic AIR model (33), which, following [32], allows to formulate the following expression for the power of band-to-band filters:

$$E \left[ \hat{h}_{k,h}(m) \right]^2 = \sum_{l=-L_w+1}^{L_w-1} \sigma^2 h \chi_h(mB+l) e^{-\frac{2(mB+l)}{\tau_h}} w^2(l). \quad (41)$$
This expression leads to an approximate relationship between the power of band-to-band filters with different frame indices, which is valid for all $L_R \in \mathbb{N}$:

$$E \left[ \bar{h}_{k,k}(m + L_R) \right]^2 = \sum_{l=-L_w+1}^{L_w-1} \sigma_h^2 \chi_h \left( (m + L_R) B + l \right) \cdot e^{-\frac{2[(m + L_R) B + l]}{\tau_h} w^2(l)} \quad \text{(42)}$$

$$\approx e^{-\frac{2L_R B}{\tau_h}} E \left[ \bar{h}_{k,k}(m) \right]^2. \quad \text{(43)}$$

If the AIR were of infinite length, recursion (43) would be exact. It now allows us to formulate a recursive expression for the expectation of the power of the STDFT of the reverberant signal $s(l)$ with respect to the AIR, denoted by $E_{h(l)} \left[ \bar{S}(m,k) \right]^2$. Employing (13) and assuming $E_{h(l)} \left[ E^{(S)}(m,k) \right]$ to be zero allows this quantity to first be expressed as

$$E_{h(l)} \left[ \bar{S}(m,k) \right]^2 = C_p \left( \sum_{m'=0}^{L_R-1} |X(m - m', k)|^2 E \left[ \bar{h}_{k,k}(m') \right]^2 \quad \text{(44)} \right.$$

$$+ \sum_{m'=L_R}^{L_H} |X(m - m', k)|^2 E \left[ \bar{h}_{k,k}(m') \right]^2 \left. \right).$$

Performing variable substitution in the second summand of (44) and subsequently using (43), we arrive at

$$E_{h(l)} \left[ \bar{S}(m,k) \right]^2 = C_p \left( \sum_{m'=0}^{L_R-1} |X(m - m', k)|^2 E \left[ \bar{h}_{k,k}(m') \right]^2 \quad \text{(45)} \right.$$

$$+ C_p \sum_{m'=0}^{L_H} |X((m - L_R) - m', k)|^2 E \left[ \bar{h}_{k,k}(m') \right]^2$$

$$\approx C_p \left( \sum_{m'=0}^{L_R-1} |X(m - m', k)|^2 E \left[ \bar{h}_{k,k}(m') \right]^2 \quad \text{(46)} \right.$$

$$+ e^{-\frac{2L_R B}{\tau_h}} E_{h(l)} \left[ \bar{S}(m - L_R,k) \right]^2 \left. \right)$$

for $1 \leq L_R \leq L_H$. Note that since the AIR is assumed to be of finite length, the implicit change of the upper limit in the second sum of (45) from $L_H - L_R$ to $L_H$ preserves the equality of the right sides of (44) and (45). A recursive expression for the MPSC of the reverberant signal may now be obtained from (46) by approximating the expectation of $\bar{h}_{k,k}(m')^2$ by (15) while neglecting the remaining expectation operator. After application of the mel filter bank, the resulting functional expression is given by

$$S_{m,q} = C_p \left( \sum_{m'=0}^{L_R-1} H_{m',q} X_{m-m',q} + e^{-\frac{2L_R B}{\tau_h}} \bar{S}_{m-L_R,q} + e^{s(R)} \quad \text{(47)} \right.$$

$$=: \hat{S}_{m,q} + e^{s(R)}. \quad \text{(48)}$$

In the additional presence of noise, the required MPSC $S_{m-L_R,q}$ is not observable. Assuming the noise MPSC to be known, a reasonable estimate for it may be obtained by taking the expectation of $\bar{S}_{m-L_R,q}$ given the noisy reverberant MPSC $Y_{m-L_R,q}$ and the MPSC of the noise $N_{m-L_R,q}$. With $Y_{m-L_R,q}$ and $N_{m-L_R,q}$ given, any remaining uncertainty about the reverberant MPSC $S_{m-L_R,q}$ is due to the phase factor $\alpha_{m-L_R,q}$. Utilizing the statistical properties of the phase factor $\alpha$ we essentially assume it to be a realization of a white, stationary and ergodic random process (see [33] for details) — the conditional expectation may thus be approximated by

$$E \left[ \bar{S}_{m-L_R,q} \right] Y_{m-L_R,q} N_{m-L_R,q} \approx Y_{m-L_R,q} - \left( 1 - 2\sigma_{\alpha}^2 \right) N_{m-L_R,q} \quad \text{(49)}$$

where $\sigma_{\alpha}^2$ denotes the variance of the zero-mean phase factor in the mel band $q$. The expression of the reverberant MPSC in the noisy reverberant case is thus given by

$$S_{m,q} = C_p \left( \sum_{m'=0}^{L_R-1} H_{m',q} X_{m-m',q} + e^{-\frac{2L_R B}{\tau_h}} \bar{S}_{m-L_R,q} \quad \text{(50)} \right.$$

$$\cdot \max \left\{ Y_{m-L_R,q} - \left( 1 - 2\sigma_{\alpha}^2 \right) N_{m-L_R,q}, 0 \right\} + e^{s(R)} \quad \text{(51)}$$

where we employed the maximum operation to ensure non-negativeness of the corresponding term. Note that $\hat{S}_{m,q} + e^{s(R)}$ may also be expressed in terms of $S_{m,q}$ and the error due to the approximation of $S_{m-L_R,q}$ as

$$\Delta S_{m-L_R,q} = S_{m-L_R,q} - \left( 1 - 2\sigma_{\alpha}^2 \right) N_{m-L_R,q} \quad \text{(52)}$$

$$= S_{m,q} + e^{s(R)} \quad \text{(53)}$$

The two approximations of the reverberant MPSCs $S_{m,q}$ by $\hat{S}_{m,q} + e^{s(R)}$ in the reverberant case and $S_{m,q} + e^{s(R)}$ in the noisy reverberant case can now directly be used instead of $S_{m,q}$ to arrive at the definitions of two vector valued observation mappings $f_{s,L_R}$ and $f_{y,L_R}$, which are appropriate for the absence and presence of noise, respectively:

$$f_{s,L_R} \left( \mathbf{x}_{m-L_R+1:m}, \mathbf{h}_0; L_R-1, y_{m-L_R} \right) \quad \text{(54)}$$

$$:= \ln \left\{ C_p \left( \sum_{m'=0}^{L_R-1} e^{x_{m-m'}+x_{m-m'}+e^{-\frac{2\sigma_{\alpha}^2}{\tau_h}} e^{s_{m-L_R}}} \quad \text{(55)} \right. \right.$$
By comparing the observation mappings (54) and (55) to (24) and (31), respectively, it can be verified that the number of exponential terms is distinctly reduced if $L_R \ll L_H$ holds. In practice, this reduction does not only lead to a reduction of the computational effort for their evaluation, but also implies a reduced amount of memory required for the storage of involved LMPSC vectors [2].

Eventually, the reverberant and noisy reverberant observation can be expressed in terms of the observation mappings (54) and (55), respectively, by

$$s_m = f_s(R) \left( x_{m-L_R+1:m}, \bar{h}_{0:L_R-1}, s_{m-L_R} \right) + v^{(s,R)}_{m,L_R}, \quad (56)$$

$$y_m = f_y(R) \left( x_{m-L_R+1:m}, \bar{h}_{0:L_R-1}, y_{m-L_R}, n_m, n_{m-L_R} \right) + v^{(y,R)}_{m,L_R}, \quad (57)$$

with corresponding error vectors $v^{(s,R)}_{m,L_R,q}$ and $v^{(y,R)}_{m,L_R,q}$. The observation error component $v^{(s,R)}_{m,L_R,q}$ is defined by

$$v^{(s,R)}_{m,L_R,q} := s_m - \bar{s}^{(s,R)}_{m,q} = \ln \left\{ \frac{S_{m,q}}{C_p \sum_{m'=0}^{L_R-1} \mathcal{H}_{m',q}^{m',m} + e^{-\frac{2L_R R}{\alpha}} S_{m-L_R,q}} \right\}. \quad (58)$$

The observation error in the presence of noise can now again be expressed in terms of the observation error in the absence of noise. The resulting functional dependency per mel band can be obtained by plugging (51) into (26) and utilizing that $\mathcal{E}^{(y,R)}_{m,L_R,q} = \mathcal{E}^{(s,R)}_{m,L_R,q} + e^{-\frac{2L_R R}{\alpha}} \Delta S_{m-L_R,q}^{v,y}$. Solving for the observation error eventually yields

$$v^{(y,R)}_{m,L_R,q} := y_m - \ln \left\{ e^{\mathcal{E}^{(s,R)}_{m,L_R,q}} + e^{n_{m,q}} \right\} = \ln \left\{ 1 + \left\{ e^{\mathcal{E}^{(s,R)}_{m,L_R,q}} e^{\Delta S_{m-L_R,q}^{v,y}} - 1 \right\} \frac{1}{1 + e^{n_{m,q} - \mathcal{E}^{(y,R)}_{m,L_R,q}}} \right\} + 2 \alpha_m e^{\mathcal{E}^{(s,R)}_{m,L_R,q}} \frac{e^{\Delta S_{m-L_R,q}^{v,y}}}{1 + e^{n_{m,q} - \mathcal{E}^{(y,R)}_{m,L_R,q}}}, \quad (60)$$

where

$$e^{\Delta S_{m-L_R,q}^{v,y}} = 1 + e^{-\frac{2L_R R}{\alpha}} \frac{\Delta S_{m-L_R,q}^{v,y}}{\bar{s}^{(s,R)}_{m,q}}. \quad (62)$$

captures the error resulting from the approximation of the reverberant MPSC $S_{m-L_R,q}$ by the conditional expectation (49) followed by the maximum operation.

Again, the observation error in the noisy reverberant case depends on the observation error $v^{(s,R)}_{m,L_R}$ in the reverberant case, the phase factor $\alpha_{m,q}$ and the difference of the noise LMPSC $n_{m,q}$ to the LMPSC $v^{(y,R)}_{m,L_R,q}$ derived from (51). This time, the recursion length $L_R$ and the approximation error $\Delta S_{m-L_R,q}$ also influence the overall observation error. However, with the recursion length $L_R$ chosen sufficiently large, $e^{\Delta S_{m-L_R,q}^{v,y}}$ will approximately be 1 and Eq. (61) becomes the recursive equivalent to (30).

IV. EXPERIMENTAL INVESTIGATIONS

After having derived several stochastic observation models theoretically, we verify their validity experimentally.

According to these models, the observed features are modeled to consist of a deterministic observation mapping function and a stochastic observation error term. Replacing the unknown AIR representation $\bar{h}_{0:L_H}$ by its mean $\bar{\mu}_{0:L_H}$, as mentioned in Sec. III-C, the observed reverberant but noise-free feature vector $s_m$ is modeled as

$$s_m = f_s \left( x_{m-L_H:m}, \bar{\mu}_{0:L_H} \right) + v^{(s)}_{m}, \quad (63)$$

$$s_m = f_s \left( x_{m-L_R+1:m}, \bar{\mu}_{0:L_R-1}, s_{m-L_R} \right) + \bar{v}^{(s,R)}_{m,L_R}, \quad (64)$$

Similarly, the observation in the noisy reverberant case is represented as

$$y_m = f_y \left( x_{m-L_H+1:m}, \bar{\mu}_{0:L_H} \right) + v^{(y)}_{m}, \quad (65)$$

$$y_m = f_y \left( x_{m-L_R+1:m}, \bar{\mu}_{0:L_R-1}, y_{m-L_R}, n_m, n_{m-L_R} \right) + \bar{v}^{(y,R)}_{m,L_R}, \quad (66)$$

both times for the non-recursive and the recursive observation model, respectively. Due to the decomposition of the observation vectors into a deterministic part and an additive stochastic part, the corresponding conditional observation PDFs, i.e.,

$$p \left( s_m \mid x_{m-L_H:m} \right), \quad (67)$$

$$p \left( s_m \mid x_{m-L_R+1:m}, s_{m-L_R} \right), \quad (68)$$

in the absence of noise and

$$p \left( y_m \mid x_{m-L_H+1:m}, n_m \right), \quad (69)$$

$$p \left( y_m \mid x_{m-L_R+1:m}, y_{m-L_R}, n_m, n_{m-L_R} \right), \quad (70)$$

in the presence of noise, are completely characterized by the conditional PDFs of the observation errors of (63)-(66). Since the observation PDFs (67)-(70) are an integral part of robust feature compensation schemes [2], [33]–[35], the following analyses take a closer look at them by means of analyzing the distribution of the observation errors. After describing the experimental setup we will address the choice of parameters of the observation mapping functions, analyze the statistical properties of the observation error and discuss the influence of the power compensation constant.

A. Experimental setup for the analysis of the observation error

For the following analysis we assume the LMPSCs of the observable signals $\bar{z}(l)$ and $\overline{f}(l)$ and those of the underlying individual signal components, i.e., the clean speech signal $\overline{f}(l)$ and the noise-only signal $\overline{n}(l)$, to be available.

Under these assumptions, estimates of the observation error in the absence and in the presence of noise may be computed. Since simultaneous recordings of the individual signal components are usually not available in practice, we used artificially generated signals for the error analysis. As clean speech signals we took the training data of the AURORA 5 database [36], which is based on the TIDigits corpus [37]...
down-sampled to a sampling rate of \( T_q^{-1} = 8 \text{ kHz} \). As noise signals, the noises used for the creation of the artificially distorted training and testing data of the AURORA 5 database were made available to us. The reverberant signals \( \Xi(l) \) were computed by convolving the clean speech signals \( \Xi(l) \) with artificially created AIRs, generated by the image method [38].

We assume the observation model to be employed for feature enhancement and for subsequent ASR or similar purposes, where the particular testing scenario is usually unknown at the training stage. Hence, the following simulation scheme has been chosen to account for this uncertainty in the AIR: We employed for the image method a cuboid measuring \( 5 \text{ m} \times 6 \text{ m} \times 3 \text{ m} \) (width \times depth \times height), which corresponds to an average-sized living room and computed 100 AIRs for three different average reverberation times, i.e., \( \hat{T}_{60} \in \{ 250 \text{ ms}, 350 \text{ ms}, 450 \text{ ms} \} \). In each turn, the individual artificial AIR is generated by randomly drawing the reverberation time from the interval \( \hat{T}_{60} = 50 \text{ ms}, \hat{T}_{60} = 1800 \text{ ms} \) in order to simulate estimation errors. Such an accuracy seems to be realistic for state-of-the-art reverberation time estimators [39].

We further randomly chose the position of the speaker to lie within one half of the room and the position of the microphone within the other half of the room. Note that the analyses following present detailed results mainly for \( \hat{T}_{60} = 450 \text{ ms} \), however, as it has already been observed in [22, Tab. VIII] and [30], we found the observation model to be insensitive to estimation errors of the reverberation time in the range of the aforementioned \( \pm 50 \text{ ms} \).

B. Parameters of the observation mapping functions

The observation models call for the evaluation of deterministic mapping functions to predict the LMPSCs of the reverberant and noisy speech signal from that of the clean speech and noise signal. Their evaluation requires, among other parameters, the expected log mel power spectral representation of the AIR, whose computation is based on the decay constant \( \tau_h \) and the energy parameter \( \sigma_h \) [22].

An estimate of the decay constant may be obtained from an estimate of the reverberation time \( T_{60} \) and the relationship (35). The energy parameter \( \sigma_h \) may be determined from the ratio of the average powers of the reverberant and the non-reverberant speech, as it is detailed in [30].

A further parameter is the length of the LMPSC representation of the AIR, which is specified by the parameter \( L_H \) for the non-recursive and by the parameter \( L_R \) for the recursive observation model.

Given an estimate of the AIR length \( L_h \), the parameter \( L_H \) can be determined from (12). A reasonable deterministic relation between the AIR length \( L_h \) and the reverberation time is given by (36). This relation was determined from the desired property that the relative power of the neglected part of the AIR resulting from the length constraint has to be less than \( \epsilon_h \). A reasonable choice is \( \epsilon_h = 10^{-3} \). The resulting AIR length \( \hat{L}_h \) and the lengths \( \hat{L}_H \) of the log mel spectral AIR representation for, e.g., \( \hat{T}_{60} = 450 \text{ ms} \) are \( \hat{L}_h = 1800 \) and \( \hat{L}_H = 24 \). It is particularly worth noting that both \( \hat{L}_h \) and \( \hat{L}_H \) grow linearly with the reverberation time, as can also be seen by plugging (35) into (36) and the resulting value into (12).

C. Properties of the observation error in the absence of background noise

In a first experiment, we analyzed the covariance matrices \( \Sigma_{v(s)} \) of the observation error in the reverberant-only case for different reverberation times \( T_{60} \). We observed, that the covariance matrices \( \Sigma_{v(s)} \) are dominated by their diagonal elements and that the correlations between the observation error components are strictly positive. Substantial elements on the secondary diagonals are mainly due to the overlap of adjacent mel bands. However, for a simplified modeling it is reasonable to assume the individual observation error components to be uncorrelated.

To render a potential feature compensation scheme analytically and thus computationally tractable it is also desirable to have a GAUSSIAN distributed observation error. Hence, we further checked whether the PDF of the observation error can be approximated by a GAUSSIAN PDF. For that purpose, we computed the normalized histograms of the true observation error components \( \hat{v}_{m,q}^{(s)} \) and compared them with the corresponding GAUSSIAN PDF approximations. The results are depicted in Fig. 2 for different mel bands for a reverberation time estimate of \( \hat{T}_{60} = 450 \text{ ms} \). They reveal the following observations.

\[
\begin{align*}
\text{Fig. 2. Normalized true \text{(H)} and \text{Gaussian \text{(G)}} \text{ approximations for different \text{mel bands} } q \text{ \text{and a reverberation time estimate} } \hat{T}_{60} = 450 \text{ ms}.}
\end{align*}
\]

The GAUSSIAN approximations match the histograms quite well and, as indicated by further experiments, become more accurate with increasing values of the reverberation time. Moreover, the variance of the observation error decreases with the mel band index \( q \), irrespective of the reverberation time.

The results obtained for the recursive observation mapping look quite similar except for one subtle difference. For lower mel bands \( q \) and low recursion lengths \( L_R \) the histograms of the observation error \( \hat{v}_{m,L_R}^{(s)} \) tend to be leptokurtotic, i.e. have more acute peaks. However, if the recursion length \( L_R \) is increased, the shape of the histograms approaches that of a GAUSSIAN PDF.

D. Properties of the observation error in the presence of background noise

In the presence of background noise, as Eq. (30) already indicates, the observation error \( v_{m,q}^{(y)} \) depends on the observation error \( v_{m,q}^{(s)} \) in the reverberant case, the phase factor \( \alpha_{m,q} \) and the difference of the noise-only LMPSC \( n_{m,q} \) to the speech-related LMPSC \( s_{m,q} \) defined in (21).
A similar relation can be found for the observation error $\tilde{v}_{m,q}^{(y)}$ of the recursive observation model.

1) The non-recursive observation model: The difference of the LMPSC $\tilde{s}_{m,q}$ to the LMPSC of the noise $n_{m,q}$ can be interpreted as a frame and mel band specific ratio of the instantaneous reverberant speech power to the noise power (instantaneous reverberant-to-noise ratio, IRNR), defined by

$$r_{m,q} := \frac{10}{\ln(10)} \left( \frac{\tilde{s}_{m,q} - n_{m,q}}{n_{m,q}} \right) \ln_{10} \left( \frac{\tilde{s}_{m,q}}{n_{m,q}} \right).$$

(71)

Thus, the observation error $\tilde{v}_{m,q}^{(y)}$ given in (30) can also be expressed in terms of this IRNR, resulting in

$$\tilde{v}_{m,q}^{(y)} \approx \ln \left\{ 1 + \left( e^{r_{m,q}} - 1 \right) \cdot \zeta (r_{m,q}) + 2 \alpha_{m,q} e^{v_{m,q}} \cdot \nu (r_{m,q}) \right\},$$

(72)

where $\zeta (r_{m,q})$ and $\nu (r_{m,q})$ are defined as

$$\zeta (r_{m,q}) := \frac{1}{1 + e^{-\frac{\ln(10)}{\ln(10)} r_{m,q}}} = \frac{1}{1 + e^{-\frac{\ln(10)}{\ln(10)} r_{m,q}}} = 1 + 2 \tanh \left( \frac{\ln(10)}{20} r_{m,q} \right),$$

(73)

$$\nu (r_{m,q}) := \frac{e^{-\frac{\ln(10)}{\ln(10)} r_{m,q}}}{1 + e^{-\frac{\ln(10)}{\ln(10)} r_{m,q}}} = \frac{1}{2} \text{sech} \left( \frac{\ln(10)}{20} r_{m,q} \right).$$

(74)

and $\tanh(\cdot)$ and $\text{sech}(\cdot)$ denote the hyperbolic tangent and secant, respectively. Both functions $\zeta (r_{m,q})$ and $\nu (r_{m,q})$ are sketched in Fig. 3 over a range of IRNRs. The values of $\zeta (r_{m,q})$ and $\nu (r_{m,q})$ at the current IRNR $r_{m,q}$ control the influence of $e^{v_{m,q}} - 1$ and $2 \alpha_{m,q} e^{v_{m,q}}$ on the resulting observation error. As a consequence, the resulting observation error highly depends on the IRNR $r_{m,q}$ and the phase factor $\alpha_{m,q}$ [40]. In particular, for $r_{m,q} \to \infty$, the observation error approaches that of the reverberant-only case and for $r_{m,q} \to -\infty$ it approaches zero. However, as will soon be shown, the PDF of the observation error may accurately be approximated by a GAUSSIAN, if its mean and variance are made time-variant, or, more precisely, variant, with respect the IRNR. Approximating the distribution of the error component $e^{v_{m,q}}$ by a GAUSSIAN distribution with IRNR-variant mean and variance calls for the computation of the mean and the variance of (72) in dependence of $r_{m,q}$. This, however, is equivalent to approximating the distribution of the exponential error component $e^{v_{m,q}}$ by a log-Normal distribution. Hence, the required statistics are $E \left[ e^{v_{m,q}} | r_{m,q} \right]$ and $\text{Var} \left[ e^{v_{m,q}} | r_{m,q} \right]$. For their computation, the phase factor $\alpha_{m,q}$ and the error of the observation model for reverberant speech $v_{m,q}$ are assumed to be mutually independent (an assumption that can reasonably well be justified). In [33] it has further been shown, that i) the phase factor is approximately independent of the IRNR and ii) that all its central moments can be computed analytically. Particularly, employing the zero-mean property of the phase factor, the required statistics are given by

$$E \left[ e^{e^{v_{m,q}}} | r_{m,q} \right] = 1 + E \left[ e^{i v_{m,q}} - 1 \right] \zeta (r_{m,q}),$$

(75)

$$\text{Var} \left[ e^{e^{v_{m,q}}} | r_{m,q} \right] = \text{Var} \left[ e^{i v_{m,q}} - 1 \right] \zeta^2 (r_{m,q}) + 4 \sigma_{\alpha_q}^2 E \left[ e^{i v_{m,q}} \right] \nu^2 (r_{m,q}).$$

(76)

Here, $\sigma_{\alpha_q}^2$ denotes the variance of the zero-mean phase factor $\alpha_{m,q}$. Since the observation error in the reverberant case $\tilde{v}_{m,q}$ is assumed to be GAUSSIAN distributed, the required moments of the (consequently log-Normally distributed) exponential error term $e^{v_{m,q}}$ can be expressed in terms of its mean $\mu_{e^{v_{m,q}}}$ and its variance $\sigma_{e^{v_{m,q}}}^2$ [41]. In particular,

$$E \left[ e^{v_{m,q}} - 1 \right] = \mu_{e^{v_{m,q}}} + \frac{1}{2} \sigma_{e^{v_{m,q}}}^2 - 1,$$

(77)

$$\text{Var} \left[ e^{v_{m,q}} - 1 \right] = \left( \sigma_{e^{v_{m,q}}}^2 - 2 \right) \left( \mu_{e^{v_{m,q}}} + \sigma_{e^{v_{m,q}}}^2 \right).$$

(78)

Finally, the mean and the variance of the GAUSSIAN approximation to the distribution of $\tilde{v}_{m,q}^{(y)}$ as a function of the IRNR $r_{m,q}$ are given in terms of the mean (75) and the variance (76) of the corresponding log-Normal approximation to the distribution of $e^{v_{m,q}}$ by

$$\mu_{\tilde{v}_{m,q}^{(y)}} (r_{m,q}) = \ln \left\{ E \left[ e^{v_{m,q}} | r_{m,q} \right] \right\} - \frac{1}{2} \sigma_{e^{v_{m,q}}}^2 (r_{m,q}),$$

(79)

$$\sigma_{\tilde{v}_{m,q}^{(y)}}^2 (r_{m,q}) = \ln \left\{ \frac{\text{Var} \left[ e^{v_{m,q}} | r_{m,q} \right]}{E \left[ e^{v_{m,q}} | r_{m,q} \right]^2} \right\}.$$  

(80)

The quality of the GAUSSIAN fit to the true distribution of the error and the importance of considering the contribution of the phase factor to it shall be illustrated by Fig. 4.

Figure 4a shows the histograms of the observation error as a function of the IRNR $r_{m,q}$ at a reverberation time $T_{60} = 450$ ms for mel band $q = 10$ and a global broadband RNR of 10 dB. The global broadband RNR thereby denotes the per-utterance ratio of the power of the reverberant speech to the power of the noise. Each row represents the distribution of the observation error $\tilde{v}_{m,q}^{(y)}$ given the instantaneous RNR $r_{m,q}$ (to be read off the y-axis). Figure 4b shows the GAUSSIAN approximation of this distribution employing (75) and (76) to obtain the mean (79) and the variance (80). The shape of both histograms and GAUSSIAN approximations consistently follow the findings of the previous discussion based on Eq. (72). As expected, it can be seen that the error variance decreases considerably as the value of $r_{m,q}$ decreases. This can best be seen by looking at the supporting solid and dashed lines, indicating the means and the unit standard deviation contours, respectively. It is worth noting that the histograms of the observation error conditioned on the IRNR were very similar, independent of the value of the global broadband RNR, which, in our experiments, showed to have only an effect on the
a priori probability of the IRNR. Finally, Fig. 4c illustrates the importance of considering the contribution of the phase factor $\alpha_{m,q}$ to the observation error. Without consideration of the phase factor, which amounts to setting $\alpha_{m,q}$ and consequently $\sigma_q^2$ to zero in (72) and (76), respectively, a severe mismatch of the GAUSSIAN approximation and the true distribution can be observed. To further examine how much of the uncertainty in the noisy reverberant observations $y_m$ is eventually removed by the deterministic observation mapping we consider the ratio $\gamma_q$ of the sample variance of the observation errors and the sample variance of the noisy reverberant observations themselves in mel band $q$. Clearly, a perfect prediction would not leave any uncertainty about the observation, resulting in a variance ratio of $\gamma_q = 0$. On the other hand, a deterministic observation mapping that is not capable of reducing the uncertainty would result in a variance ratio $\gamma_q = 1$. Figure 5 shows the variance ratio for reverberation times of $T_{60} = 350$ ms and $T_{60} = 450$ ms at a global broadband RNR of 10 dB. The variance ratio is clearly below 1 for all mel bands. Further, one finds it to be lower at $T_{60} = 350$ ms than at $T_{60} = 450$ ms. This quite intuitive finding can be explained as follows: for lower reverberation times, less non-causal and less cross terms in (9) eventually make up the error term in (13), which eventually contributes to the observation error. Hence, predicting the observation at $T_{60} = 350$ ms can be carried out more certain than at $T_{60} = 450$ ms.

2) The recursive observation model: The observation error $v_{m,L_R,q}$ in the recursive observation model is given in (60). The difference of the LMPSC $s_{m,q}^{(y,R)}$ to the LMPSC of the noise $r_{m,q}$ can again be interpreted as a frame and mel band specific ratio of the instantaneous reverberant speech power to the noise power. However, this time it is defined by

$$v_{m,L_R,q} = \ln \{ 10 \} \left( \frac{s_{m,q}^{(y,R)}}{s_{m,q}} - 1 \right) = 10 \log_{10} \left( \frac{s_{m,q}^{(y,R)}}{s_{m,q}} \right) .$$

Thus, the observation error $v_{m,L_R,q}$ given in (61) can also be expressed in terms of the IRNR $r_{m,q}$, resulting in

$$v_{m,L_R,q} = \ln \left\{ 1 + \left( e^{r_{m,L_R,q}} - 1 \right) \delta_{m,L_R,q} \right\}$$

$$+ 2 \alpha_{m,q} e^{r_{m,q}} - \nu_{m,q} \right\},$$

where $\delta$ and $\nu$ are defined in (73) and (74), respectively, and are evaluated at $r_{m,q}$ instead of $r_{m,q}$.

Fig. 6 shows the histograms of the observation error $v_{m,L,q}$ for different recursion lengths $L_R$ as a function of the IRNR $r_{m,q}$. For ease of comparison with Fig. 4, the histograms are again given at $T_{60} = 450$ ms, mel band $q = 10$ and a global broadband RNR of 10 dB.

As expected, the uncertainty in the estimation of $S_{m,L,R,q}$ cause the histograms to significantly differ from the ones obtained from the non-recursive observation model. In particular, the histograms exhibit negative skewness’ at mid and low levels of the IRNR and the variances of the observation error are larger than those of the error in the non-recursive observation model. However, with an increasing recursion length $L_R$, the influence of the estimation error becomes negligible and the histograms of the observation error approach those of the non-recursive observation model. Hence, for $L_R$ chosen sufficiently large, the derived GAUSSIAN approximation may reasonably be applied again.

E. Relevance of power compensation constant

In a last experiment we analyzed the sensitivity of our observation model with respect to the power compensation constant $C_I$. We consider the non-recursive observation model in the absence and the presence of background noise. Denoting the optimal value of the power compensation constant by
C_{P}^{\text{opt}}$ (see Sec. III-D for its computation), the actual value of the power compensation constant may be expressed in terms of $C_{P}^{\text{opt}}$ by $C_{P} = \xi C_{P}^{\text{opt}}$. The parameter $\xi$ thus specifies the deviation of the chosen power compensation constant $C_{P}$ from the optimal value $C_{P}^{\text{opt}}$. Looking at the non-recursive observation model, the observation error in the absence of noise defined in (23) can now be written as

$$
\tilde{v}_{m,q}^{(s)} = \psi_{m,q}^{(s,\text{opt})} \ln \{ \xi \},
$$

where $\psi_{m,q}^{(s,\text{opt})}$ is the error corresponding to the optimal power compensation constant. Hence, any deviation from the optimal power compensation constant only affects the mean of the observation error $\tilde{v}_{m,q}^{(s)}$.

In the presence of noise, the observation error $v_{m,q}^{(s)}$ can be reformulated in terms of $\psi_{m,q}^{(s,\text{opt})}$ and $\xi$ to give

$$
v_{m,q}^{(s)} = \ln \left\{ 1 + \left( e^{\psi_{m,q}^{(s,\text{opt})}} - 1 \right) \frac{1}{\xi} \right\} - \ln \{ \xi \} + \frac{1}{\sqrt{\xi}} \psi_{m,q}^{(s,\text{opt})} \nu \left( \ln \{ \xi \} \right) + 2 \alpha_{m,q} \frac{1}{\sqrt{\xi}} \psi_{m,q}^{(s,\text{opt})} \nu \left( \ln \{ \xi \} \right) .
$$

Note that we used the IRNR defined in (71) for the optimal power compensation constant to actually see the influence of a sub-optimally chosen compensation constant on the observation error. From (84) it can be seen that a deviation from $C_{P}^{\text{opt}}$ will, due to the non-linearity, affect all moments of the error distribution and also shift the distribution with respect to the IRNRs. Nevertheless, the observation error may again be approximated by a GAUSSIAN distribution with IRNR-variant mean and variance, as we have done before (compare (79) and (80)). To illustrate this behavior, we consider the histograms and the GAUSSIAN distributions of the observation error $v_{m,q}^{(y)}$ for $C_{P} = 1$, i.e., $\xi = 1/C_{P}^{\text{opt}}$. The results are presented in Fig. 7. For low values of the instantaneous RNR the sub-

![Fig. 6. The histograms of the observation error $\tilde{v}_{m,q}^{(y)}$ in the recursive observation model as a function of the IRNR $v_{m,q}^{(R)}$ at $T_{60} = 450$ ms, mel band $q = 10$ and a global broadband RNR of $10$ dB for different recursion lengths $L_{R}$. The supporting solid and dashed lines indicate the means and the unit standard deviation contours, respectively.](image)

![Fig. 7. The histograms (a) of the observation error $\tilde{v}_{m,q}^{(y)}$ in the non-recursive observation model and the GAUSSIAN approximations (b) for $C_{P} = 1$ as a function of the IRNR $v_{m,q}$ at $T_{60} = 450$ ms, mel band $q = 10$ and a global broadband RNR of $10$ dB. The supporting solid and dashed lines indicate the means and the unit standard deviation contours, respectively.](image)

V. CONCLUSION

In this paper we have developed stochastic observation models describing the relationship between the logarithmic mel power spectral coefficients of clean speech, noise and noisy reverberant speech. Different observation mapping functions have been developed, which, by introducing a power compensation constant, allow for an unbiased prediction of the noisy reverberant features from clean and noise-only features.

Special emphasis has been placed on analyzing and modeling the observation error, i.e., the difference between the true, observed features and their prediction by the models. The experiments revealed that additive noise significantly changes the statistics of the error: while in the absence of noise the error can be considered to be a realization of a white, stationary and ergodic GAUSSIAN stochastic process, noise leads to a severe dependency of the observation error on the instantaneous ratio of reverberant power to noise power (IRNR). However, a GAUSSIAN approximation of the observation error can still be found by considering the mean and the variance as a function of the IRNR.

The findings presented in this paper not only shed more light on the complicated interaction between speech, noise and noisy reverberant speech features, they may also be
of significant practical importance. The observation models developed here can for instance be used within a model-based feature compensation framework for the robust recognition of noisy reverberant speech, as described in [24].

ACKNOWLEDGMENT

The authors would like to thank Hans-Günter Hirsch for providing the AIRs as well as the interior noise recordings used for the compilation of the AURORA 5 database.

REFERENCES


