© 2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Nonnegative HMM for Babble Noise Derived from Speech HMM: Application to Speech Enhancement

Nasser Mohammadiha*, *Student Member, IEEE,* and Arne Leijon, *Member, IEEE*

Abstract

Deriving a good model for multitalker babble noise can facilitate different speech processing algorithms, e.g. noise reduction, to reduce the so-called cocktail party difficulty. In the available systems, the fact that the babble waveform is generated as a sum of N different speech waveforms is not exploited explicitly. In this paper, first we develop a gamma hidden Markov model for power spectra of the speech signal, and then formulate it as a sparse nonnegative matrix factorization (NMF). Second, the sparse NMF is extended by relaxing the sparsity constraint, and a novel model for babble noise (gamma nonnegative HMM) is proposed in which the babble basis matrix is the same as the speech basis matrix, and only the activation factors (weights) of the basis vectors are different for the two signals over time. Finally, a noise reduction algorithm is proposed using the derived speech and babble models. All of the stationary model parameters are estimated using the expectation-maximization (EM) algorithm, whereas the time-varying parameters, i.e. the gain parameters of speech and babble signals, are estimated using a recursive EM algorithm. The objective and subjective listening evaluations show that the proposed babble model and the final noise reduction algorithm significantly outperform the conventional methods.

Index Terms

Babble noise, hidden Markov model, nonnegative matrix factorization, speech enhancement.

I. Introduction

Multispeaker babble noise is one of the frequently encountered interferences in daily life that greatly degrades the quality and intelligibility of a target speech signal. The problem of understanding the

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

The authors are with the Sound and Image Processing Laboratory, School of Electrical Engineering, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden (e-mail: nmoh@kth.se; leijon@kth.se, phone: +46-8 790 7556).

desired speech in the presence of other interfering speech signals and background noise (also known as the "cocktail party problem") has received great attention since it was popularized by Cherry in 1953 [1]. Different auditory aspects of this problem are investigated (e.g. [2], [3]), and the intelligibility of speech in the presence of multitalker babble noise is examined (e.g. [4]). In addition, there have been few studies that have addressed some babble-specific signal processing techniques to improve speech perception in the presence of background babble noise. In [5], considering a single-channel observation of the babble noise, a framework was proposed to characterize the underlying babble signal. Also, the effect of the number of conversations and speakers was investigated, and a system was proposed to identify the number of speakers in a presented babble noise; moreover, it was shown that this information is beneficial for speaker recognition.

On the contrary, little attention has been paid to develop mathematical babble-specific models that can also be used in signal processing algorithms, e.g. speech enhancement. The goal of speech enhancement algorithms is to improve the quality and intelligibility of the noisy speech, e.g., [6]–[11], and among different applications, it is very beneficial for hearing aid users [12]. Various classes of single channel model-based speech enhancement approaches have been proposed in the literature. In these methods, for each type of signal (speech or noise) a model is considered and the model parameters are obtained using the training samples of that signal. Then, the task of the speech enhancement is done by defining an interactive model between the speech and noise signals. Some examples of this class of algorithms include the codebook-based approaches [13] and HMM-based methods [14]–[16]. However, none of these methods exploit the fact that the babble is generated by adding different speech signals, and hence the structure of the considered model for babble noise is similar to that of other noise types. In this paper, we derive a statistical model for babble noise, which takes into account the fact that the babble is generated by adding speech signals of M independent speakers. Then, we propose a single-channel speech enhancement framework that utilizes the derived babble model to enhance the noisy speech signal.

The proposed babble model is based on the nonnegative matrix factorization (NMF). NMF is a technique to approximate a nonnegative matrix $\mathbf{X}$ by a nonnegative linear combination of some basis vectors [17], i.e. $\mathbf{X} \approx \mathbf{TV}$. In speech processing: $\mathbf{X}$ is the spectrogram of the signal with short-time spectral vectors stored as columns in $\mathbf{X}$, $\mathbf{T}$ is the basis matrix or basis spectral vectors, and $\mathbf{V}$ is called the NMF coefficient matrix. NMF has been used successfully in different fields including blind source separation [18]–[20], and speech enhancement [21]–[25]. The "pure addition" property of NMF makes it a powerful technique to be used whenever some nonnegative quanta are added to each other. In the case of babble noise, spectral vectors of different speech signals can be added to generate a spectral vector

of babble.

The basic idea of NMF-based speech enhancement algorithms is that, for each signal, an NMF model is considered and its parameters are obtained using the training data. Then, a mixing model is defined, which usually involves the assumption that the spectrograms of the noise and speech signals are additive, and speech enhancement is carried out by a Wiener-type filtering approach. Two important shortcomings of NMF have to be considered when designing NMF-based speech enhancement systems:

1) The correlation between consecutive time-frames is not handled directly in a standard NMF. To overcome this problem, several approaches have been proposed [21]–[25]. For instance, a semi-supervised approach (where the noise type is not known a priori) was proposed in [22], which was based on a nonnegative hidden Markov model (NHMM) where the correlation of the signals were taken into account by the transition probability matrix of the underlying HMM. In [25], a Bayesian NMF based speech enhancement algorithm was proposed in which the temporal correlation of the underlying speech and noise signals was exploited through the informative prior distributions.

2) For some noise signals, the noise basis matrix is quite similar to the basis matrix of the speech signal, e.g. the basis matrix of the babble noise should be quite similar to the basis matrix of the speech signal. As a result, the performance of the noise reduction algorithms is usually worse in the case of babble noise [21], [24]. This issue has not been addressed in the available systems and is one of the main focuses of this study.

In this paper, first we derive an ergodic gamma-HMM model for the power spectral coefficients of the speech signal. Next, we formulate the speech model as a sparse NMF. Then, by relaxing the sparsity constraint, we derive a gamma nonnegative hidden Markov model (gamma-NHMM) for babble noise in which the basis matrix is identical to the speech basis matrix, and only the activity of the basis vectors segregates the speech from the babble signal. Moreover, an expectation-maximization (EM) algorithm is proposed to estimate the model parameters. In addition, to employ the derived babble model for speech enhancement, an HMM-based speech enhancement framework in the time-frequency domain is proposed where each power spectral vector of the power spectrogram of speech and babble signals are modeled by the gamma-HMM and gamma-NHMM models, respectively. The proposed framework differs from the state-of-the-art HMM-based approaches [14]–[16] as we directly model the spectral vectors with HMM. In the available HMM-based methods, the waveform signal is modeled as an autoregressive (AR) process, and hence the waveforms of speech and noise signals are modeled by HMM. Thus, this new framework facilitates a new class of HMM-based speech enhancement algorithms. Similar to [16], the interaction model for the noisy speech signal is constructed by considering a prior distribution over the long-term

energy levels of the speech and noise signals. A recursive EM algorithm [26], [27] is developed to estimate the time-varying parameters of these distributions online. The excellence of the proposed babble model and noise reduction scheme is demonstrated through objective evaluations and a subjective listening test.

The rest of the paper is organized as follows: The gamma-HMM speech signal model is developed in Section II. In Section III, the gamma-NHMM model of babble noise is derived. In Section IV, the mixed signal model and noise reduction algorithm is constructed. The estimation of the stationary model parameters and time-varying parameters is described in Section V. The objective and subjective examination of the noise reduction algorithms are presented in Section VI. Finally, Section VII concludes the study.

II. Speech Signal Model

A. Single-voice Gamma HMM

We model the magnitude-squared DFT coefficients (periodogram coefficients) of the speech signal using an $\bar{N}$-state HMM with gamma distributions as output probability density functions. Throughout this paper, random variables are represented with capital letters, e.g. $\mathbf{X} = [X_{kt}]$ denotes the matrix of random variables associated with the DFT coefficients of the clean speech, where k is the frequency bin and t denotes the time-frame index. The corresponding realizations are shown with small letters, e.g. $\mathbf{x}$=$[x_{kt}]$. Also, let $|\cdot|^2$ represents the element-wise magnitude-square operator. The conditional distribution of $|X_{kt}|^2$ is given as:

$$f\left(|x_{kt}|^2 \mid \bar{S}_t = i, G_t = g_t\right) = \frac{\left(|x_{kt}|^2\right)^{\alpha_k - 1}}{(g_t b_{ki})^{\alpha_k}\,\Gamma(\alpha_k)} e^{-|x_{kt}|^2/(g_t b_{ki})}, \tag{1}$$

where the conditional density $f_{X|Y}(x \mid Y = y)$ is simply shown as $f(x \mid Y = y)$ to keep notations uncluttered, and $\Gamma(\cdot)$ is the Gamma function. Here, $\bar{S}_t$ is the random hidden state of the speech signal, α_k is the shape parameter, b_{ki} is the scale parameter, and G_t is the stochastic gain parameter, which is discussed later. The expected value and variance of X_{kt} are defined as: $E(|X_{kt}|^2 \mid \bar{S}_t = i, G_t = g_t) = \alpha_k g_t b_{ki}$, and var$(|X_{kt}|^2 \mid \bar{S}_t = i, G_t = g_t) = \alpha_k (g_t b_{ki})^2$.

The gamma assumption for a magnitude-squared DFT coefficient in (1) is motivated by the super-Gaussianity of the speech DFT coefficients [9], [11]. Denote the real and imaginary parts of the DFT coefficient by Re$\{X_{kt}\}$ and Im$\{X_{kt}\}$, respectively. Assuming that Re$\{X_{kt}\}$ and Im$\{X_{kt}\}$ have a two-sided generalized gamma distribution is equivalent to assuming that $|$Re$\{X_{kt}\}|$ and $|$Im$\{X_{kt}\}|$ have a generalized gamma distribution. Then, it can be easily shown that $|$Re$\{X_{kt}\}|^2$ and $|$Im$\{X_{kt}\}|^2$ have a

gamma distribution if the γ parameter of the generalized gamma distributions equals 2 (see [11] for a general discussion and definition of γ). We use the standard assumption that $\mathrm{Re}\{X_{kt}\}$ and $\mathrm{Im}\{X_{kt}\}$ are independent and identically distributed. Since the sum of two independent gamma random variables (RV) with equal scale parameters is a gamma RV, $|X_{kt}|^2 = |\mathrm{Re}\{X_{kt}\}|^2 + |\mathrm{Im}\{X_{kt}\}|^2$ will have a gamma distribution.

In general, state-conditional densities can describe different parts of the speech signal depending on the total number of states. For example, when 50$\sim$60 states are available, each state roughly corresponds to one phoneme.

The short-term stochastic gain parameter G_t in (1) is considered to model the long-term changes in the speech energy level over time. Since G_t is nonnegative, we choose to have a gamma distribution to govern G_t in order to simplify the resulting algorithm:

$$f\left(g_t\right) = \frac{g_t^{\phi-1}}{\theta_t^{\phi}\Gamma\left(\phi\right)} e^{-g_t/\theta_t}, \tag{2}$$

where ϕ and θ_t are the shape and scale parameters, respectively. In this model, the long-term speech level is modeled by the time-varying scale parameter θ_t, while relative signal-energy levels for different states are modeled by b_{ki} (see Fig. 1). Also, we have: $E\left(G_t\right) = \phi\theta_t$, and $\mathrm{var}\left(G_t\right) = \phi\theta_t^2$. Since $\sqrt{\mathrm{var}\left(G_t\right)}/E\left(G_t\right) = \sqrt{\phi}$, by using (2) we assume that in the log-domain the standard deviation of outcomes of G_t from its mean value is approximately constant, independent of the long-term level of speech. Considering that $E\left(G_t\right)$ is updated for different speech levels, the above assumption of gamma distribution for G_t is reasonable.

The complete HMM output density functions can now be expressed as:

$$f\Big(|\mathbf{x}_t|^2 \mid \bar{S}_t = i, G_t = g_t\Big) = \prod_{k=1}^{K} f\Big(|x_{kt}|^2 \mid \bar{S}_t = i, G_t = g_t\Big), \tag{3}$$

where we have assumed that DFT coefficients at different frequency bins are conditionally independent [6], [9], [11]. The state-conditional probability of the observed power spectral coefficients of the speech signal (which will be used for parameter estimation) can now be computed by integrating out the gain variable. Using the properties of the generalized inverse Gaussian distribution (see Appendix B), this can be obtained in a closed form as:

$$\begin{aligned} f\Big(|\mathbf{x}_t|^2 \mid \bar{S}_t = i\Big) &= \int_0^{\infty} f\Big(|\mathbf{x}_t|^2 \mid \bar{S}_t = i, G_t = g_t\Big) f\left(g_t\right) dg_t \\ &= \frac{2\tau^{\nu/2}\mathcal{K}_{\vartheta}\left(2\sqrt{\rho\tau}\right)}{\rho^{\nu/2}\theta_t^{\phi}\Gamma\left(\phi\right)} \prod_{k=1}^{K} \frac{|x_{kt}|^{2(\alpha_k-1)}}{b_{ki}^{\alpha_k}\Gamma\left(\alpha_k\right)}, \end{aligned}$$

where we have defined $\rho = 1/\theta_t$, $\nu = \phi - \sum_{k=1}^{K}\alpha_k$, $\tau = \sum_{k=1}^{K}|x_{kt}|^2 b_{ki}^{-1}$, and $\mathcal{K}_{\vartheta}\left(\cdot\right)$ denotes a modified Bessel function of the second kind.

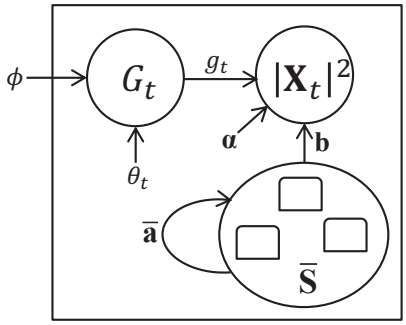


Fig. 1: A schematic representation of the HMM with gain modeling. $\bar{\mathbf{a}}$ dentes the transition probability matrix.

The sequence of hidden states is characterized by a first-order Markov model, with the transition probability matrix $\bar{\mathbf{a}}$, and with the elements $\bar{a}_{ij} = f(\bar{S}_t = j \mid \bar{S}_{t-1} = i)$. As we are modeling speech in general, and not a specific utterance, the state Markov chain is considered to be fully connected, and hence *ergodic*, with the time-invariant state probability mass vector $\bar{\mathbf{p}}$, and with the elements $\bar{p}_i = f(\bar{S}_t = i)$.

B. Gamma-HMM as a Probabilistic NMF

Instead of denoting the hidden state by its index number, as $\bar{S}_t = i$, we can denote the random discrete state by a one-of-$\bar{N}$ indicator column vector $\bar{\mathbf{S}}_t$, where $\bar{S}_{it} = 1$ and $\bar{S}_{jt} = 0, j \neq i$. Using this notation, the selected state-conditional set of scale parameters $\mathbf{b}_i = (b_{1i}, \ldots, b_{Ki})^\top$, given a particular state $\bar{\mathbf{S}}_t$ with $\bar{S}_{it} = 1$, can be simply expressed by $\mathbf{b}\bar{\mathbf{S}}_\mathbf{t}$, where all of the $\bar{N}$ state-conditional scale parameter vectors have been collected as columns in the "basis" matrix $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{\bar{N}})$.

The complete sequence of scale parameter vectors for the complete spectrum sequence $|\mathbf{X}|^2 = (|\mathbf{X}_1|^2, |\mathbf{X}_2|^2, \ldots, |\mathbf{X}$ can then be expressed as $\mathbf{b}\bar{\mathbf{S}}$ where $\bar{\mathbf{S}} = (\bar{\mathbf{S}}_1, \bar{\mathbf{S}}_2, \ldots, \bar{\mathbf{S}}_t, \ldots)$ is the random sequence of the state indicator vectors. The probability of the complete sequence $|\mathbf{x}|^2$ of the observed short-time spectra, given any state sequence $\bar{\mathbf{s}}$ and gain factors $\mathbf{g} = (g_1, g_2, \ldots, g_t, \ldots)$, can now be obtained as:

$$f\left(|\mathbf{x}|^2 \mid \bar{\mathbf{s}}, \mathbf{g}\right) = \prod_t f\left(|\mathbf{x}_t|^2 \mid \bar{\mathbf{s}}_t, g_t\right), \tag{4}$$

$$f\Big(|\mathbf{x}_t|^2 \mid \bar{\mathbf{s}}_t, g_t\Big) = \prod_k \frac{\left(|x_{kt}|^2\right)^{\alpha_k - 1}}{\left(g_t \left[\mathbf{b}\bar{\mathbf{s}}_t\right]_k\right)^{\alpha_k} \Gamma\left(\alpha_k\right)} e^{-|x_{kt}|^2 / \left(g_t [\mathbf{b}\bar{\mathbf{s}}_t]_k\right)}, \tag{5}$$

where $[\cdot]_k$ denotes the k^{th} element of the vector, and we have: $E(|X_{kt}|^2 \mid \bar{\mathbf{s}}_t, g_t) = \alpha_k g_t \left[\mathbf{b}\bar{\mathbf{s}}_t\right]_k$. Eq. (5) can be used to derive an NMF representation of any observed nonnegative matrix $|\mathbf{x}|^2$. In order to show this, we approximate an observed sequence by its expected value, under the model assumptions, and show that the expected value is the product of two nonnegative matrices. To compute the expected value,

the posterior distribution of the state and gain variables are employed. That is, an NMF approximation $\widehat{|\mathbf{x}_t|^2}$ of an input vector $|\mathbf{x}_t|^2$ is given as:

$$\widehat{|\mathbf{x}_t|^2} = \sum_{\bar{\mathbf{s}}_t} \int E\left(|\mathbf{X}_t|^2 \mid \bar{\mathbf{s}}_t, g_t\right) f\left(\bar{\mathbf{s}}_t, g_t \mid |\mathbf{x}|^2\right) dg_t. \tag{6}$$

Let us define $\hat{\mathbf{b}}$ with elements $\hat{b}_{ki} = \alpha_k b_{ki}$. Noting from (5) that $E(|\mathbf{X}_t|^2 \mid \bar{\mathbf{s}}_t, g_t) = g_t \hat{\mathbf{b}} \bar{\mathbf{s}}_t$, (6) can be written as:

$$\widehat{|\mathbf{x}_t|^2} = \hat{\mathbf{b}} \sum_{\bar{\mathbf{s}}_t} \int g_t \bar{\mathbf{s}}_t f\left(\bar{\mathbf{s}}_t \mid |\mathbf{x}|^2\right) f\left(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}|^2\right) dg_t. \tag{7}$$

Here, the conditional state probabilities $f(\bar{\mathbf{s}}_t \mid |\mathbf{x}|^2)$ can be calculated using the forward-backward algorithm [28]. Since g_t depends only on the current observation, $f(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}|^2) = f(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}_t|^2)$. The posterior distribution of the gain variable is a generalized inverse Gaussian distribution (this is derived in Appendix B and will also be used in Subsection V-A). Thus, the required integration in (7) is available in a closed form (Eq. (48)). Denoting $E(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}_t|^2) = \int g_t f(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}_t|^2) dg_t$, and $\mathbf{u}_t = \sum_{\bar{\mathbf{s}}_t} \bar{\mathbf{s}}_t f(\bar{\mathbf{s}}_t \mid |\mathbf{x}|^2) E(g_t \mid \bar{\mathbf{s}}_t, |\mathbf{x}_t|^2)$, we can write: $\widehat{|\mathbf{x}_t|^2} = \hat{\mathbf{b}}\mathbf{u}_t$. Hence, the proposed gamma-HMM model can be used to factorize a nonnegative matrix $|\mathbf{x}|^2$ into a nonnegative basis matrix $\hat{\mathbf{b}}$ and an NMF coefficients matrix $\mathbf{u}$ as: $|\mathbf{x}|^2 \approx \hat{\mathbf{b}}\mathbf{u}$. In an extremely sparse case where $f(\bar{\mathbf{s}}'_t \mid |\mathbf{x}|^2) = 1$ only for one state $\bar{\mathbf{s}}'_t$, depending on time t, and all of the other states have a zero probability, this model reduces to: $|\mathbf{x}|^2 \approx \hat{\mathbf{b}}\mathbf{u}'$ with $\mathbf{u}'_t = \bar{\mathbf{s}}'_t E(g_t \mid \bar{\mathbf{s}}'_t, |\mathbf{x}_t|^2)$.

III. Probabilistic Model of Babble Noise

We model the waveform of the babble noise as a weighted sum of M i.i.d. clean speech sources. Therefore, the expected value of the short-time power spectrum vector (periodogram) of babble at time t, $|\mathbf{V}_t|^2$, is given by:

$$E\left(|\mathbf{V}_t|^2\right) = \sum_{m=1}^{M} E\left(|\mathbf{X}_{mt}|^2\right), \tag{8}$$

where each random vector $|\mathbf{X}_{mt}|^2$ is independently generated by an instance of the gamma-HMM described in Section II. Note that in (8) different weights are used for different speakers as a consequence of the gain modeling in Section II. That is, there is a hidden speaker-dependent gain (g_{mt}) in (8) (see also (1)). Eq (8) provides a simplified model of real-life babble noise because we are not modeling reverberations here. There might also be additional noise with a recorded babble, which is not considered in (8). However, it must be mentioned that all of the babble model parameters will be estimated given a

babble training data set, with no information about M. Therefore, the estimated parameters will be such that the model explains the considered babble as well as possible.

In this view, the babble noise is still described by an HMM with discrete states defined by the combination of the states for each of the M sources. Since the speech signal has $\bar{N}$ states, there are $\bar{N}^M$ possible discrete states for the babble. As the discrete HMM for the clean speech is already an approximation, and speech should probably rather be modeled with a continuous-state HMM, it would be preferable to describe the babble sequence with a continuous-state HMM. On the other hand, an exact implementation of the EM algorithm for HMMs with a continuous-state is generally not possible, except for some very few specific cases, e.g. Gaussian linear state-space models, and simulation-based methods have to be used instead [29]. Hence, it would be preferable to avoid a continuous-state structure whenever it is possible. Furthermore, in a real babble noise only a finite number of states (say representative states) would be sufficient for practical purposes to model the normalized spectral shape of the signal; this is indicated, for example, by the success of the vector quantization techniques to quantize continuous signals with a limited number of centroid vectors effectively. Based on these reasons, in the following we model the babble noise with a discrete-state HMM.

Using the text following (5), (8) can be written as:

$$\begin{aligned} E\left(|\mathbf{V}_t|^2 \mid \bar{\mathbf{s}}_{1t}, g_{1t}, \ldots \bar{\mathbf{s}}_{Mt}, g_{Mt}\right) = \\ \textstyle\sum_{m=1}^{M} E\left(|\mathbf{X}_{mt}|^2 \mid \bar{\mathbf{s}}_{mt}, g_{mt}\right) = \\ \textstyle\sum_{m=1}^{M} g_{mt}\hat{\mathbf{b}}\bar{\mathbf{s}}_{mt} = \hat{\mathbf{b}} \textstyle\sum_{m=1}^{M} g_{mt}\bar{\mathbf{s}}_{mt}, \end{aligned} \tag{9}$$

where $\hat{b}_{ki} = \alpha_k b_{ki}$ as before. In the babble HMM, we will now approximate the sum over m in (9) by the babble hidden state vectors and the gain variables. Let us denote the babble hidden state vector at time t by $\ddot{\mathbf{S}}_t$ (as opposed to the speech state indicator shown by $\bar{\mathbf{S}}_t$) and its realizations by $\ddot{\mathbf{s}}_t$ that can take one of the $\ddot{N}$ possible state value vectors $\{\ddot{\mathbf{s}}'_1, \ddot{\mathbf{s}}'_2, \ldots \ddot{\mathbf{s}}'_{\ddot{N}}\}$. Note that $\ddot{N}$ is different from the number of speakers in the babble, which is shown by M in (9). Also, denote the stochastic babble gain by random variable H_t and its realizations by h_t.

The power spectrum values of the babble, as defined by (8), (9) are not exactly gamma-distributed[1], given the hidden state. However, our informal simulations showed that the babble DFT coefficients also have super-Gaussian distributions. This is understandable, considering the similarity of speech and babble.

[1]Eq. (9) is defined for the expected values; to obtain the exact distribution of the babble power spectral vectors, given the hidden states for all of the speakers, both the summation of individual gamma distributions and the distribution of the cross terms have to be considered.

Hence, the same argument that was used for the speech model in Subsection II-A can be used here to motivate that gamma distribution is a good approximation for the distribution of the babble spectra. We now extend the clean-speech model in (5), just slightly, to model the density of the babble short-time power spectrum as:

$$f\Big(|\mathbf{v}_t|^2 \mid \ddot{\mathbf{s}}_t, h_t\Big) = \prod_k \frac{\Big(|v_{kt}|^2\Big)^{\beta_k - 1}}{\left(h_t \left[\mathbf{b}\ddot{\mathbf{s}}_t\right]_k\right)^{\beta_k} \Gamma\left(\beta_k\right)} e^{-|v_{kt}|^2 / \left(h_t [\mathbf{b}\ddot{\mathbf{s}}_t]_k\right)}, \tag{10}$$

here, the main new feature is that the hidden state vectors $\ddot{\mathbf{s}}$ are not indicator vectors (columns of $\bar{\mathbf{s}}$ were one-of-$\bar{N}$ indicator vectors in (5)). This is a result of (8). More specifically, if we set $\beta_k = \alpha_k$ and $\ddot{\mathbf{s}}_t = \sum_m g_{mt} \bar{\mathbf{s}}_{mt}$, then (10) leads to the same expected value as in (9) with $h_t = 1$. In this context, $\ddot{\mathbf{s}}_t$ is the weighted sum of the M indicator vectors. The shape parameters β_k are still assumed to be independent of the hidden states, but may be different from the shape parameters α_k of the clean speech model. In this approach, the babble signal is generated as a weighted sum of different clean speech waveforms, thus, the "basis" matrix $\mathbf{b}$ is assumed to be the same and only the weights of the basis vectors are different for the speech and the babble signals. The short-term stochastic gain H_t in (10) is assumed to have a gamma distribution as:

$$f\left(h_t\right) = \frac{h_t^{\psi - 1}}{\gamma_t^{\psi} \Gamma\left(\psi\right)} e^{-h_t / \gamma_t}. \tag{11}$$

In (11), the scale parameter γ_t represents the long-term energy level of the babble signal, and ψ is the shape parameter. An EM-based algorithm is proposed in Subsection V-B to estimate $\ddot{N}$ babble state value vectors, $\ddot{\mathbf{s}}'_i$ for $i = 1, \ldots \ddot{N}$, the state transition probabilities $\ddot{a}_{ij} = f(\ddot{\mathbf{S}}_t = \ddot{\mathbf{s}}'_j \mid \ddot{\mathbf{S}}_{t-1} = \ddot{\mathbf{s}}'_i), \beta_k, \psi$, and γ_t given the recorded babble noise. Eq. (10) is referred as gamma-NHMM since the described model performs an NMF on the scale parameters of the HMM output distributions, which are gamma distributions.

IV. Speech Enhancement Method

A noise reduction scheme is proposed in this section to enhance the speech signal that is degraded by the babble noise. The mixed signal model is described in Subsection IV-A, which is used later in Subsection IV-B to derive an MMSE estimator for the speech signal.

In the proposed models, the power spectra of the clean speech and the babble noise are conditionally gamma-distributed. Even though a gamma distribution might be a good approximation for the power spectra of the mixed signal, obtaining the MMSE estimator for the clean speech signal is difficult for this case ([30] proposes a solution uing this approximation). Therefore, in this part of the paper (which provides an application of the developed babble model) we limit the models to use exponential

distributions for the speech and the babble power spectral coefficients ($\alpha_k = 1$ in (1), $\beta_k = 1$ in (10)). This corresponds to the assumption that speech and babble DFT coefficients have complex Gaussian distributions, which have been used successfully in the literature (e.g. [6], [31]). In the experimental section, we show that even with this additional simplification the proposed noise reduction method outperforms the competing algorithms. To keep the generality of the speech and babble models for potential future applications, the proposed parameter estimation algorithm in Section V will be given for the general gamma case.

A. Clean Speech Mixed with Babble

Assuming that the DFT coefficients of the clean speech and babble noise are complex Gaussian, DFT coefficients of the mixed signal $\mathbf{Y}$,

$$\mathbf{Y}_t = \mathbf{X}_t + \mathbf{V}_t,$$

will also have complex Gaussian distribution. Let us represent the composite state of the mixed signal by $\mathbf{S}_t$ that can take one of the $N = \bar{N}\ddot{N}$ possible outcomes. Defining $\sigma^2_{Y_{kt}} = E(|Y_{kt}|^2 \mid \mathbf{s}_t, g_t, h_t) = E(|X_{kt}|^2 \mid \bar{\mathbf{s}}_t, g_t) + E(|V_{kt}|^2 \mid \ddot{\mathbf{s}}_t, h_t)$ we have:

$$f\left(y_{kt} \mid g_t, h_t, \mathbf{s}_t\right) = \frac{1}{\pi\sigma^2_{Y_{kt}}} e^{-\frac{|y_{kt}|^2}{\sigma^2_{Y_{kt}}}}, \tag{12}$$

and also

$$f\left(\mathbf{y}_t \mid g_t, h_t, \mathbf{s}_t\right) = \prod_{k=1}^{K} f\left(y_{kt} \mid g_t, h_t, \mathbf{s}_t\right). \tag{13}$$

The state-conditional distribution of the mixed signal can be obtained by integrating out the gain variables as:

$$f\left(\mathbf{y}_t \mid \mathbf{s}_t\right) = \int\int f\left(\mathbf{y}_t \mid g_t, h_t, \mathbf{s}_t\right) f\left(g_t, h_t\right) dg_t dh_t. \tag{14}$$

The required expectations to calculate $\sigma^2_{Y_{kt}}$ in (12) are obtained considering the models given in (5) and (10). The analytical evaluation of (14) turns out to be difficult; although numerical methods can be used to calculate the required integrations, we approximate the integrand by its behavior near to its maximum by applying Laplace approximation [32, Sec. 4.4]. Hence, we first derive an EM algorithm to obtain the state-dependent Maximum a-Posteriori estimates g'_t and h'_t in Appendix A based on the following optimization problem:

$$\left\{g'_t, h'_t\right\} = \underset{g_t, h_t}{\arg\max}\; f\left(\mathbf{y}_t \mid g_t, h_t, \mathbf{s}_t\right) f\left(g_t, h_t\right), \tag{15}$$

then (14) is approximated by

$$f\left(\mathbf{y}_t \mid \mathbf{s}_t\right) \approx f\left(\mathbf{y}_t \mid g'_t, h'_t, \mathbf{s}_t\right) f\left(g'_t, h'_t\right) \frac{2\pi}{\sqrt{\det\left(A_{\mathbf{s}_t}\right)}}, \tag{16}$$

where $\det(A_{\mathbf{s}_t})$ is the determinant of the negative Hessian of $\ln f(\mathbf{y}_t, g_t, h_t \mid \mathbf{s}_t)$ with respect to g_t, h_t, evaluated at the maximum point. The expression for the Hessian matrix is also given in Appendix A. It should also be mentioned that in [16] an EM algorithm was developed to find the mode of the joint distribution and then $f(y \mid \mathbf{s}_t)$ was approximated by $f(y, g'_t, h'_t \mid \mathbf{s}_t)$.

B. Clean Speech Estimator

The posterior distribution of the clean speech DFT coefficients given the noisy observations can be written as [14], [16]:

$$f\left(\mathbf{x}_t \mid \mathbf{y}_1^t\right) = \frac{\sum_{\mathbf{s}_t} \eta_t(\mathbf{s}_t) f(\mathbf{x}_t, \mathbf{y}_t \mid \mathbf{s}_t)}{f\left(\mathbf{y}_t \mid \mathbf{y}_1^{t-1}\right)}, \tag{17}$$

where $\mathbf{y}_{t1}^{t2}=\{\mathbf{y}_{t1}, \mathbf{y}_{t1+1}, \ldots, \mathbf{y}_{t2}\}$, and $\eta_t(\mathbf{s}_t)=f\left(\mathbf{s}_t \mid \mathbf{y}_1^{t-1}\right)$ is the probability of being in the composite state $\mathbf{s}_t$ at time t given all of the noisy observations until time $t-1$, and is calculated as:

$$\eta_t(\mathbf{s}_t) = f\left(\mathbf{s}_t \mid \mathbf{y}_1^{t-1}\right) = \sum_{\mathbf{s}_{t-1}} a_{\mathbf{s}_{t-1},\mathbf{s}_t} f\left(\mathbf{s}_{t-1} \mid \mathbf{y}_1^{t-1}\right), \tag{18}$$

with $a_{\mathbf{s}_{t-1},\mathbf{s}_t}=f(\mathbf{S}_t=\mathbf{s}_t \mid \mathbf{S}_{t-1}=\mathbf{s}_{t-1}) = \bar{a}_{\bar{\mathbf{s}}_{t-1},\bar{\mathbf{s}}_t} \ddot{a}_{\ddot{\mathbf{s}}_{t-1},\ddot{\mathbf{s}}_t}$ because of the independency of the speech and the noise Markov chains, and $f(\mathbf{s}_{t-1} \mid \mathbf{y}_1^{t-1})$ is the scaled forward variable obtained using the forward algorithm [28]. The joint distribution of $\mathbf{X}_t$ and $\mathbf{Y}_t$ can also be written as:

$$\begin{aligned} f(\mathbf{x}_t, \mathbf{y}_t \mid \mathbf{s}_t) &= \int\int f(\mathbf{x}_t, \mathbf{y}_t, g_t, h_t \mid \mathbf{s}_t)\, dg_t dh_t \approx \\ & f\left(\mathbf{x}_t \mid \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t\right) \int\int f(\mathbf{y}_t, g_t, h_t \mid \mathbf{s}_t)\, dg_t dh_t \approx \\ & f\left(\mathbf{x}_t \mid \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t\right) f\left(\mathbf{y}_t, g'_t, h'_t \mid \mathbf{s}_t\right) \frac{2\pi}{\sqrt{\det(A_{\mathbf{s}_t})}}, \end{aligned} \tag{19}$$

where the second line is obtained using a point approximation for $f(\mathbf{x}_t \mid \mathbf{y}_t, g_t, h_t, \mathbf{s}_t)$ (similarly to [16]), and the last line is obtained by using approximation (16). We can also write:

$$\begin{aligned} f\left(\mathbf{y}_t \mid \mathbf{y}_1^{t-1}\right) &= \sum_{\mathbf{s}_t} \int \eta_t(\mathbf{s}_t) f(\mathbf{x}_t, \mathbf{y}_t \mid \mathbf{s}_t)\, d\mathbf{x}_t \\ &\approx \sum_{\mathbf{s}_t} \eta_t(\mathbf{s}_t) f\left(\mathbf{y}_t, g'_t, h'_t \mid \mathbf{s}_t\right) \frac{2\pi}{\sqrt{\det(A_{\mathbf{s}_t})}}. \end{aligned} \tag{20}$$

Denoting $\zeta_t(\mathbf{s}_t, \mathbf{y}_t) = \eta_t(\mathbf{s}_t) f(\mathbf{y}_t, g'_t, h'_t \mid \mathbf{s}_t) \frac{2\pi}{\sqrt{\det(A_{\mathbf{s}_t})}}$, and using (19) and (20), (17) can be written as:

$$f\left(\mathbf{x}_t \mid \mathbf{y}_1^t\right) = \frac{\sum_{\mathbf{s}_t} \zeta_t(\mathbf{s}_t, \mathbf{y}_t) f\left(\mathbf{x}_t \mid \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t\right)}{\sum_{\mathbf{s}_t} \zeta_t(\mathbf{s}_t, \mathbf{y}_t)}. \tag{21}$$

Because of the Gaussian assumption, calculating the state-conditional posterior distribution of the clean speech DFT coefficients is straightforward and is given by a complex Gaussian distribution with the mean value obtained via the Wiener filtering:

$$E\left(X_{kt} \mid \mathbf{y}_t, g'_t, h'_t, \mathbf{s}_t\right) = C_{X_{kt}} \left(C_{X_{kt}} + C_{V_{kt}}\right)^{-1} y_{kt}, \tag{22}$$

and the covariance matrix given as:

$$E\left(\left|X_{kt}-E\left(X_{kt}\mid \mathbf{y}_t,g'_t,h'_t,\mathbf{s}_t\right)\right|^2\mid \mathbf{y}_t,g'_t,h'_t,\mathbf{s}_t\right)$$
$$=C_{X_{kt}}-C_{X_{kt}}\left(C_{X_{kt}}+C_{V_{kt}}\right)^{-1}C_{X_{kt}}, \tag{23}$$

where $C_{X_{kt}}=E(|X_{kt}|^2\mid g'_t,\mathbf{s}_t)=\alpha_k g'_t\left[\mathbf{b}\bar{\mathbf{s}}_t\right]_k$ and $C_{V_{kt}}=E(|V_{kt}|^2\mid h'_t,\mathbf{s}_t)=\beta_k h'_t\left[\mathbf{b}\ddot{\mathbf{s}}_t\right]_k$. By using (22) and (21), the MMSE estimator of the clean speech DFT coefficients is derived as:

$$\begin{aligned}\hat{\mathbf{x}}_t &= E\left(\mathbf{X}_t\mid \mathbf{y}_1^t\right)\\ &= \frac{\sum_{\mathbf{s}_t}\zeta_t\left(\mathbf{s}_t,\mathbf{y}_t\right)E\left(\mathbf{X}_t\mid \mathbf{y}_t,g'_t,h'_t,\mathbf{s}_t\right)}{\sum_{\mathbf{s}_t}\zeta_t\left(\mathbf{s}_t,\mathbf{y}_t\right)},\end{aligned} \tag{24}$$

or equivalently as $\hat{x}_{kt}=\kappa_{kt}y_{kt}$ where the gain parameter is given by:

$$\kappa_{kt}=\frac{\sum_{\mathbf{s}_t}\zeta_t\left(\mathbf{s}_t,\mathbf{y}_t\right)C_{X_{kt}}\left(C_{X_{kt}}+C_{V_{kt}}\right)^{-1}}{\sum_{\mathbf{s}_t}\zeta_t\left(\mathbf{s}_t,\mathbf{y}_t\right)}. \tag{25}$$

V. Parameter Estimation

A. Speech Model Training

The EM-based Baum-Welch algorithm is followed to train speech and noise models [28], [33]. The parameters of the speech model are denoted by $\lambda=\{\bar{\mathbf{a}},\mathbf{b},\boldsymbol{\alpha},\phi,\theta\}$. Letting the training data consist of R speech utterances, it is assumed that the time-dependent scale parameter of the stochastic gain, θ, remains constant during each utterance for simplicity, hence, denoted by θ_r in the following.

Denote the whole training set by $\bar{\mathbf{o}}=\{(\bar{\mathbf{o}}_{1,1}\ \ldots\bar{\mathbf{o}}_{1,T_1}),\ldots(\bar{\mathbf{o}}_{R,1}\ \ldots\bar{\mathbf{o}}_{R,T_R})\}$ where $\bar{\mathbf{o}}_{r,t}=[\bar{o}_{r,kt}]$ represents the speech power spectral vector of the r^{th} sentence at time t. Similarly, let $\bar{\mathbf{Z}}=\{\bar{\mathbf{S}},\mathbf{G}\}$ represent the hidden variables in the model, which are not observed. Then, the maximization step in the EM algorithm consists of maximizing

$$Q\left(\hat{\lambda},\lambda\right)=\sum_{\bar{\mathbf{s}}}\int f\left(\bar{\mathbf{z}}\mid\bar{\mathbf{o}},\lambda\right)\ln\left(f\left(\bar{\mathbf{z}},\bar{\mathbf{o}}\mid\hat{\lambda}\right)\right)d\mathbf{g}, \tag{26}$$

w.r.t $\hat{\lambda}$, where λ is the estimated parameters from the previous iteration of the EM algorithm. $Q(\hat{\lambda},\lambda)$ can be written as:

$$Q\left(\hat{\lambda},\lambda\right)=\hat{Q}\left(\hat{\lambda},\lambda\right)+\sum_{r,t,i}\omega_{t,r}\left(i\right)\int f\left(g_{r,t}\mid\bar{\mathbf{o}}_{r,t},\bar{S}_{r,t}=i,\lambda\right)\cdot$$
$$\left(\ln f\left(g_{r,t}\mid\hat{\lambda}\right)+\ln f\left(\bar{\mathbf{o}}_{r,t}\mid g_{r,t},\bar{S}_{r,t}=i,\hat{\lambda}\right)\right)dg_{r,t}, \tag{27}$$

for $r=1\ldots R$, $t=1\ldots T_r$, and $i=1\ldots\bar{N}$. Here, $\hat{Q}(\hat{\lambda},\lambda)$ includes the terms for optimizing the transition matrix $\hat{\mathbf{a}}$ and is maximized using the standard Baum-Welch algorithm. The posterior state probabilities

$$\omega_{t,r}\left(i\right)=f\left(\bar{S}_{r,t}=i\mid\bar{\mathbf{o}},\lambda\right),$$

are obtained by the forward-backward algorithm [28]. To obtain the new parameters, (27) is differentiated w.r.t the parameters of interest, and the result is set to zero. The objective function in (27) is separable for on the one hand $\{\hat{\mathbf{b}}, \hat{\boldsymbol{\alpha}}\}$, and on the other hand $\{\hat{\phi}, \hat{\theta}\}$. First, consider $\{\hat{\mathbf{b}}, \hat{\boldsymbol{\alpha}}\}$; obtaining the gradient w.r.t. $\hat{b}_{ki}$ and setting it to zero yields the following estimate:

$$\hat{b}_{ki} = \frac{\sum_{r,t} \omega_{t,r}(i)\, \bar{o}_{r,kt} E_{G_{r,t}|\bar{S}_{r,t},\lambda}\left(G_{r,t}^{-1}\right)}{\hat{\alpha}_k \sum_{r,t} \omega_{t,r}(i)} = \frac{\mu_{ki}^o}{\hat{\alpha}_k}, \tag{28}$$

where $E_{G_{r,t}|\bar{S}_{r,t},\lambda}(\cdot)$ is the expectation w.r.t. the posterior distribution of the gain variable $f\left(g_{r,t} \mid \bar{\mathbf{o}}_{r,t}, \bar{S}_{r,t} = i, \lambda\right)$, and μ_{ki}^o is defined as:

$$\mu_{ki}^o = \frac{\sum_{r,t} \omega_{t,r}(i)\, \bar{o}_{r,kt} E_{G_{r,t}|\bar{S}_{r,t},\lambda}\left(G_{r,t}^{-1}\right)}{\sum_{r,t} \omega_{t,r}(i)}. \tag{29}$$

Inserting (28) into (27), and setting the gradient of the objective function w.r.t. $\hat{\alpha}_k$ to zero yields:

$$\varphi(\hat{\alpha}_k) - \ln(\hat{\alpha}_k) = \frac{1}{\sum_{r,t,i} \omega_{t,r}(i)} \times$$
$$\sum_{r,t,i} \omega_{t,r}(i) \left(\ln \bar{o}_{r,kt} - E_{G_{r,t}|\bar{S}_{r,t},\lambda}(\ln G_{r,t}) - \ln \mu_{ki}^o\right), \tag{30}$$

where $\varphi(u) = \frac{d}{du} \ln \Gamma(u)$ is the digamma function, and μ_{ki}^o is defined in (29). Hence, (30) is solved first, e.g. by Newton's method, and the obtained $\hat{\alpha}_k$ is inserted into (28) to estimate $\hat{b}_{ki}$. Similarly, $\hat{\phi}$ and $\hat{\theta}$ can be obtained by first estimating the shape parameter ϕ as:

$$\varphi\left(\hat{\phi}\right) - \ln\left(\hat{\phi}\right) = \frac{1}{\sum_{t,r,i} \omega_{t,r}(i)} \times$$
$$\sum_{t,r,i} \omega_{t,r}(i) \left(E_{G_{r,t}|\bar{S}_{r,t},\lambda}(\ln G_{r,t}) - \ln \mu_r^g\right), \tag{31}$$

with μ_r^g defined as:

$$\mu_r^g = \frac{\sum_{t,i} \omega_{t,r}(i)\, E_{G_{r,t}|\bar{S}_{r,t},\lambda}(G_{r,t})}{\sum_{t,i} \omega_{t,r}(i)},$$

and then using $\hat{\phi}$ to obtain $\hat{\theta}_r = \mu_r^g / \hat{\phi}$. Since the gamma probability density function given in (1) is log concave in α_k and b_{ki} around the stationary points $\hat{b}_{ki}, \hat{\alpha}_k$, these update rules are guaranteed to increase the overall log likelihood score of the parameters [34]. To perform the updates, it is required to calculate the posterior expected values of the functions of the gain variables. The posterior distribution of the gain variable, $f\left(g_{r,t} \mid \bar{\mathbf{o}}_{r,t}, \bar{S}_{r,t} = i, \lambda\right)$, is obtained in Appendix B and is a generalized inverse Gaussian. The required expected values $E_{G_{r,t}|\bar{S}_{r,t},\lambda}(G_{r,t})$, $E_{G_{r,t}|\bar{S}_{r,t},\lambda}\left(G_{r,t}^{-1}\right)$, and $E_{G_{r,t}|\bar{S}_{r,t},\lambda}(\ln G_{r,t})$ are given with (48), (49), and (50), respectively.

B. Babble Model Training

The parameters of the babble model are denoted by $\lambda = \{\ddot{\mathbf{a}}, \ddot{\mathbf{s}}', \boldsymbol{\beta}, \psi, \gamma\}$ where $\ddot{\mathbf{s}}' = \{\ddot{\mathbf{s}}'_1, \ddot{\mathbf{s}}'_2, \ldots \ddot{\mathbf{s}}'_{\ddot{N}}\}$ is the set of the $\ddot{N}$ babble state value vectors. These vectors are in principle the weighting factors

associated with the basis matrix $\mathbf{b}$. Letting the training data consist of R different recordings of babble noise, similar to the speech model, it is assumed that the time-dependent scale parameter of the stochastic gain H remains constant during each recording, hence, denoted by γ_r in the following.

Denote the whole training set by $\ddot{\mathbf{o}}$. The main difference between noise training and speech training is that for the babble training we must also update the babble state value vectors $\ddot{\mathbf{s}}'_i$ for $i = 1, \dots \ddot{N}$ simultaneously with the other parameters. The update rules for $\ddot{\mathbf{a}}, \gamma$, and ψ are similar to the update rules of $\bar{\mathbf{a}}, \theta$, and ϕ, respectively. The estimation of $\boldsymbol{\beta}$ and $\ddot{\mathbf{s}}'$ are coupled. Hence, it is easier to optimize the EM help function $Q(\hat{\lambda}, \lambda)$ w.r.t. these parameters separately, given the previous estimates of them. Obtaining the derivative of $Q(\hat{\lambda}, \lambda)$ w.r.t. β_k and setting it to zero yields the following estimate for β_k:

$$\varphi\left(\hat{\beta}_k\right) = \frac{\sum_{r,t,i} \omega_{t,r}\left(\ddot{\mathbf{s}}'_i\right)}{\sum_{r,t,i} \omega_{t,r}\left(\ddot{\mathbf{s}}'_i\right)} \times$$

$$\left(\ln \ddot{o}_{r,kt} - \ln\left[\mathbf{b}\ddot{\mathbf{s}}'_i\right]_k - E_{H_{r,t}|\ddot{\mathbf{s}}_{r,t},\lambda}\left(\ln H_{r,t}\right)\right). \tag{32}$$

The update rule for $\ddot{\mathbf{s}}'_i$ cannot be obtained in a closed form. Here, we present an approach based on the concave-convex procedure (CCCP) [35], [36] to iteratively maximize the EM help function $Q(\hat{\lambda}, \lambda)$. CCCP is a procedure to find a local minimum of a nonconvex function and is often used to minimize a cost function that can be written as a difference of convex functions. The core idea of this procedure is that the nonconvex function is approximated by a convex function, which can be easily minimized, and then the procedure is iterated until a local minimum is found. The negative EM help function for the babble model is given as:

$$-Q\left(\hat{\lambda}, \lambda\right) = Q' +$$

$$\sum_{t,r,i} \omega_{t,r}\left(\ddot{\mathbf{s}}'_i\right) \sum_k \left(\frac{\ddot{o}_{r,kt} E_{H_{r,t}|\ddot{\mathbf{s}}_{r,t},\lambda}\left(H_{r,t}^{-1}\right)}{\left[\mathbf{b}\widehat{\ddot{\mathbf{s}}}'_i\right]_k} + \hat{\beta}_k \ln\left[\mathbf{b}\widehat{\ddot{\mathbf{s}}}'_i\right]_k \right)$$

$$= Q' + \sum_i \left(P_1\left(\widehat{\ddot{\mathbf{s}}}'_i\right) + P_2\left(\widehat{\ddot{\mathbf{s}}}'_i\right)\right), \tag{33}$$

where Q' is independent of the state variables $\ddot{\mathbf{s}}'$. Due to the summation, it is optimal to minimize (33) w.r.t. each $\widehat{\ddot{\mathbf{s}}}'_i$ independently. It can be easily shown that P_1 is a convex function where P_2 is a concave function. Using the CCCP procedure, a convex problem is generated as:

$$\widehat{\ddot{\mathbf{s}}}'_i(l+1) = \arg\min_{\mathbf{x}} \; P_1(\mathbf{x}) + \mathbf{x}^\top \nabla P_2\left(\widehat{\ddot{\mathbf{s}}}'_i(l)\right), \tag{34}$$

$$s.t. \qquad \mathbf{x} \geq 0$$

where $\nabla P_2(\mathbf{s})$ represents the gradient of P_2 evaluated at $\mathbf{s}$, and l is the iteration number. This constrained problem can be solved by usual convex optimization tools, e.g. the interior-point methods [37, ch.11].

This procedure is followed iteratively until a locally optimal solution $\widehat{\ddot{\mathbf{s}}}'_i$ is obtained. Even though the CCCP procedure does not lead to a closed form solution, it makes the solution faster and more robust. The required derivatives to solve the above problem are given in Appendix C. In summary, the following algorithm is pursued to find the optimal babble state value vectors:

1) Initialize $\widehat{\ddot{\mathbf{s}}}'_i(1)$ using the solution obtained in the previous iteration of the EM algorithm for $i \in \{1, \dots \ddot{N}\}$, set $l = 1$.
2) For each i, iterate between the following steps until convergence (usually 2–3 iterations are enough):
   a) Calculate $\nabla P_2(\widehat{\ddot{\mathbf{s}}}'_i(l))$ (51).
   b) Solve problem (34) using the interior-point methods.
   c) $l = l + 1$.

Since the parameter estimation framework is based on the EM, initialization of the algorithm is important. To assign the initial values for $\ddot{\mathbf{s}}'_i$ (before the first iteration of the EM), we generated two minutes of 10-person babble noise using the TIMIT database (2f+2m + $1\text{f}_{-1.25\text{dB}}$+$1\text{f}_{-3\text{dB}}$+$1f_{-6\text{dB}}$+$1\text{m}_{-1.25\text{dB}}$+$1\text{m}_{-3\text{dB}}$+$1\text{m}_{-6\text{dB}}$), and the gamma-HMM (Subsection II-B) was applied independently to each speaker's spectrogram to find the NMF weighting vector $\mathbf{u}_t$. Then, all of the ten $\mathbf{u}_t$ vectors were summed together to obtain $\mathbf{u}_t^{\text{babble}}$. At the end, a K-means clustering procedure was applied to cluster all of the columns of $\mathbf{u}^{\text{babble}}$ into $\ddot{N}$ groups whose mean values were used to initialize the state value vectors ($\ddot{\mathbf{s}}'_i$) for the babble training.

C. Updating Time-varying Parameters

The scale parameters of the stochastic gains are time-variant and, thus, for the purposes of noise reduction they have to be estimated online given only the noisy signal. In the following, the parameters $\lambda_t = \{\theta_t, \gamma_t\}$ are updated in a recursive manner after the estimation of the clean speech signal. Therefore, given the noisy signal, a correction term is calculated and is added to the current estimates to obtain the new estimate of the parameters. In the remainder of this section, this correction term is obtained and is used to update the time-varying parameters (Eq. (40) and (41)). An algorithm was presented in [27] to estimate the HMM parameters online, that was based on the recursive EM algorithm and the stochastic approximation [26], [38]. Here, we follow a similar procedure as described in [16], [27] to update λ_t. The recursive EM algorithm is a stochastic approximation in which the parameters of interest are updated sequentially. To do so, the EM help function is defined as the conditional expectation of the log likelihood of the complete data until the current time w.r.t. posterior distribution of the hidden variables. Then, this help function is maximized over the parameters by a single-iteration stochastic approximation in each time instance. Denote the hidden random variables of the EM algorithm as $\mathbf{Z}_t = \{\mathbf{S}_t, G_t, H_t\}$. Given a

noisy observation at time t, $\mathbf{y}_t$, a new estimate of the parameters $\hat{\lambda}_t$ is obtained by solving:

$$\hat{\lambda}_t = \arg\max_{\lambda_t} Q_t\left(\lambda_t, \hat{\lambda}_1^{t-1}\right), \qquad \text{where}$$

$$Q_t\left(\lambda_t, \hat{\lambda}_1^{t-1}\right) = \sum_{\mathbf{s}_1^t} \int\int f\left(\mathbf{z}_1^t \mid \mathbf{y}_1^t, \hat{\lambda}_1^{t-1}\right) \times$$

$$\ln f\left(\mathbf{y}_1^t, \mathbf{z}_1^t \mid \lambda_t\right) d\mathbf{g}_1^t d\mathbf{h}_1^t, \tag{35}$$

where $\hat{\lambda}_1^{t-1} = \left\{\hat{\lambda}_1, \ldots \hat{\lambda}_{t-1}\right\}$, $\mathbf{z}_1^t = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_t\}$. As it is shown in [27], the objective function in (35) can be simplified, up to an additive constant, as:

$$Q_t\left(\lambda_t, \hat{\lambda}_1^{t-1}\right) = \text{const.} + \sum_{\tau=1}^{t} \mathcal{L}_{\tau|t}\left(\lambda_t, \hat{\lambda}_1^{\tau-1}\right),$$

with

$$\mathcal{L}_{\tau|t}\left(\lambda_t, \hat{\lambda}_1^{\tau-1}\right) = \sum_{\mathbf{s}_\tau} \int\int f\left(\mathbf{z}_\tau \mid \mathbf{y}_1^t, \hat{\lambda}_1^{\tau-1}\right) \times$$

$$\left(\ln f\left(g_\tau \mid \mathbf{s}_\tau, \lambda_t\right) + \ln f\left(h_\tau \mid \mathbf{s}_\tau, \lambda_t\right)\right) dg_\tau dh_\tau. \tag{36}$$

The parameters of interest can be updated as [27]:

$$\hat{\lambda}_t = \hat{\lambda}_{t-1} + \left(-\frac{\partial^2 Q_t\left(\lambda_t, \hat{\lambda}_1^{t-1}\right)}{\partial \lambda_t^2}\right)^{-1} \left.\frac{\partial \mathcal{L}_{t|t}\left(\lambda_t, \hat{\lambda}_1^{t-1}\right)}{\partial \lambda_t}\right|_{\hat{\lambda}_{t-1}}. \tag{37}$$

Using the Bayes rule, the posterior probability of the hidden states can be written as:

$$f\left(\mathbf{z}_\tau \mid \mathbf{y}_1^t, \hat{\lambda}_1^{\tau-1}\right) = \frac{f\left(\mathbf{z}_\tau, \mathbf{y}_\tau \mid \mathbf{y}_1^{\tau-1}, \mathbf{y}_{\tau+1}^t, \hat{\lambda}_1^{\tau-1}\right)}{f\left(\mathbf{y}_\tau \mid \mathbf{y}_1^{\tau-1}, \mathbf{y}_{\tau+1}^t, \hat{\lambda}_1^{\tau-1}\right)} =$$

$$\frac{f\left(\mathbf{s}_\tau \mid \mathbf{y}_1^{\tau-1}, \mathbf{y}_{\tau+1}^t, \hat{\lambda}_1^{\tau-1}\right) f\left(\mathbf{y}_\tau, h_\tau, g_\tau \mid \mathbf{s}_\tau, \hat{\lambda}_1^{\tau-1}\right)}{f\left(\mathbf{y}_\tau \mid \mathbf{y}_1^{\tau-1}, \mathbf{y}_{\tau+1}^t, \hat{\lambda}_1^{\tau-1}\right)}, \tag{38}$$

where the standard Markov chain property (independency of the observations given the hidden states) is used to get the second line. To reduce the computation effort, (38) is approximated as (this is also done implicitly in [16]):

$$f\left(\mathbf{z}_\tau \mid \mathbf{y}_1^t, \hat{\lambda}_1^{\tau-1}\right) \approx$$

$$\frac{f\left(\mathbf{s}_\tau \mid \mathbf{y}_1^{\tau-1}, \hat{\lambda}_1^{\tau-1}\right) f\left(\mathbf{y}_\tau, h_\tau, g_\tau \mid \mathbf{s}_\tau, \hat{\lambda}_1^{\tau-1}\right)}{f\left(\mathbf{y}_\tau \mid \mathbf{y}_1^{\tau-1}, \hat{\lambda}_1^{\tau-1}\right)}. \tag{39}$$

For $t = \tau$, the above approximation is exact. Using (16), (20), and (39) in (36) yields:

$$\mathcal{L}_{\tau|t}\left(\lambda_t, \hat{\lambda}_1^{\tau-1}\right) \approx \sum_{\mathbf{s}_\tau} \omega_\tau\left(\mathbf{s}_\tau, \mathbf{y}_\tau\right) \times$$

$$\left(\ln f\left(g'_\tau \mid \mathbf{s}_\tau, \lambda_t\right) + \ln f\left(h'_\tau \mid \mathbf{s}_\tau, \lambda_t\right)\right),$$

where $\omega_\tau\left(\mathbf{s}_\tau,\mathbf{y}_\tau\right)=\frac{\zeta_\tau(\mathbf{s}_\tau,\mathbf{y}_\tau)}{\sum_{\mathbf{s}_\tau}\zeta_\tau(\mathbf{s}_\tau,\mathbf{y}_\tau)}$ is the scaled forward variable, and $\zeta_t\left(\mathbf{s}_t,\mathbf{y}_t\right)=\eta_t\left(\mathbf{s}_t\right)f\left(\mathbf{y}_t,g_t',h_t'\mid\mathbf{s}_t\right)\frac{2\pi}{\sqrt{\det(A_{\mathbf{s}_t})}}$ as in Subsection IV-B. Evaluating (37) for θ_t and γ_t yields:

$$\hat{\theta}_t=\hat{\theta}_{t-1}+\frac{1}{\mathcal{I}_t\left(\hat{\theta}_{t-1}\right)}\sum_{\mathbf{s}_t}\omega_t\left(\mathbf{s}_t,\mathbf{y}_t\right)\left(\frac{-\phi}{\hat{\theta}_{t-1}}+\frac{g_t'}{\hat{\theta}_{t-1}^2}\right), \tag{40}$$

$$\mathcal{I}_t\left(\hat{\theta}_{t-1}\right)=\sum_{\tau=1}^{t}\sum_{\mathbf{s}_\tau}\omega_\tau\left(\mathbf{s}_\tau,\mathbf{y}_\tau\right)\left(\frac{-\phi}{\hat{\theta}_{t-1}^2}+\frac{2g_t'}{\hat{\theta}_{t-1}^3}\right).$$

and

$$\hat{\gamma}_t=\hat{\gamma}_{t-1}+\frac{1}{\mathcal{I}_t\left(\hat{\gamma}_{t-1}\right)}\sum_{\mathbf{s}_t}\omega_t\left(\mathbf{s}_t,\mathbf{y}_t\right)\left(\frac{-\psi}{\hat{\gamma}_{t-1}}+\frac{h_t'}{\hat{\gamma}_{t-1}^2}\right), \tag{41}$$

$$\mathcal{I}_t\left(\hat{\gamma}_{t-1}\right)=\sum_{\tau=1}^{t}\sum_{\mathbf{s}_\tau}\omega_\tau\left(\mathbf{s}_\tau,\mathbf{y}_\tau\right)\left(\frac{-\psi}{\hat{\gamma}_{t-1}^2}+\frac{2h_t'}{\hat{\gamma}_{t-1}^3}\right).$$

To ensure the required positivity of the step sizes $1/\mathcal{I}_t(\hat{\theta}_{t-1})$ in (40) and $1/\mathcal{I}_t(\hat{\gamma}_{t-1})$ in (41) [27], [38], and to take care of the time-variant parameters, we can modify $\mathcal{I}_t(\hat{\lambda}_{t-1})$ slightly by adding a restriction and forgetting factors to reduce the effect of the previous observations as [16], [27]:

$$\begin{aligned}\mathcal{I}_t\left(\hat{\theta}_{t-1}\right) &= \xi_\theta\mathcal{I}_{t-1}\left(\hat{\theta}_{t-2}\right)+ \\ &\quad \max\left(\beta_\theta,\sum_{\mathbf{s}_t}\omega_t\left(\mathbf{s}_t,\mathbf{y}_t\right)\left(\frac{-\phi}{\hat{\theta}_{t-1}^2}+\frac{2g_t'}{\hat{\theta}_{t-1}^3}\right)\right),\end{aligned} \tag{42}$$

$$\begin{aligned}\mathcal{I}_t\left(\hat{\gamma}_{t-1}\right) &= \xi_\gamma\mathcal{I}_{t-1}\left(\hat{\gamma}_{t-2}\right)+ \\ &\quad \max\left(\beta_\gamma,\sum_{\mathbf{s}_t}\omega_t\left(\mathbf{s}_t,\mathbf{y}_t\right)\left(\frac{-\psi}{\hat{\gamma}_{t-1}^2}+\frac{2h_t'}{\hat{\gamma}_{t-1}^3}\right)\right),\end{aligned} \tag{43}$$

with $0<\xi_\theta,\xi_\gamma<1$, and $0<\beta_\theta,\beta_\gamma$.

VI. Experiments and Results

The capability of the proposed models and the performance of the developed noise reduction algorithm is investigated in various ways. In Subsection VI-A, the details of the implementation of the proposed system is explained. In Subsection VI-B, the developed noise reduction scheme is evaluated and compared to state-of-the-art methods using different objective measures and a subjective listening test. The performance of the developed system is compared to that of the Bayesian NMF (BNMF) based approach [25] and the ETSI (European Telecommunications Standards Institute) front end Wiener filtering [39].

In the BNMF approach, to utilize the temporal correlation of the underlying speech and noise signals, the posterior distributions of the NMF coefficients at the past time instances were widened and applied as the new prior distribution through the Bayesian framework to obtain an MMSE estimator for the speech

signal [25]. A comparison to the BNMF method has been motivated by the analogy of the proposed babble model and nonnegative matrix factorization. Also, as it is reported in [25], the BNMF-based noise reduction approach outperforms different competing algorithms. On the other hand, the ETSI two-stage Wiener filter is carefully tuned for good performance in denoising speech [39], and it is considered here to compare the performance of the model-based approaches to a standard approach that does not benefit from trained noise-specific models.

A. System Implementation

The proposed models for speech and babble signals were trained using the TIMIT and NOISEX-92 databases, respectively. All of the signals were down-sampled to 16-kHz and the DFT was implemented using a frame length of 320 samples with 50% overlapped windows using a Hann window. For speech, 600 sentences from the training set of the TIMIT were used as training data, and for babble noise the first 75% of the signal was used for training while the rest of the signal was used for the test purposes. To investigate the performance of the algorithms as a function of the number of speakers in the babble noise, a different set of babble training and testing data was used, which is explained in Subsection VI-B3. Also, the core test set of the TIMIT database (192 sentences) was exploited for the noise reduction evaluation. The signal synthesis was performed using the overlap-and-add procedure.

For the speech model, $\bar{N} = 55$ states were trained in order to roughly identify these states by different phonemes. For the babble model, a discrete HMM with $\ddot{N}$ states was considered. As a result, the final mixed signal model includes $N = \bar{N}\ddot{N}$ states. To carry out the speech enhancement and calculate the final speech gain κ_{kt} (25), the weighted sum of the state-conditional Wiener filters has to be calculated while for each of the N states, the MAP estimate of the stochastic gain variables has to be performed, which is time consuming in general. Although a large value for $\ddot{N}$ may approximate the underlying continuous state-space of the babble model better, it will result in a computationally more expensive system and for $\ddot{N}$ larger than 50 a pruning algorithm [15], [16] has to be implemented to keep the level of complexity practical. In our experiments, we set $\ddot{N} = 10$ (except Subsection VI-B3) since the performance was quite similar for $\ddot{N}$ in the range of 10 to 200. Moreover, we observed that a high shape parameter (5$\sim$30) for the stochastic gain variables makes the MAP estimation faster while the performance remains similar. Hence, in our simulations the shape parameters of both the stochastic gain variables were set to 15 although the data-driven estimate of the shape parameter of the speech stochastic gain variable was less than one (this is an indication of a high variation in the state-conditioned energy level of the signal).

For this setup, our Matlab implementation runs in approximately 5-times real time[2] using a PC with 3.8 GHz Intel CPU and 2 GB RAM. The online parameter estimation (42,43) was done using $\xi_\theta = 0.99$, $\xi_\gamma = 0.98$, and $\beta_\theta = \beta_\gamma = 100$, which were set experimentally.

Additionally, motivated by our previous work [24], an exponential smoothing was performed as $\bar{\kappa}_{kt} = 0.4\bar{\kappa}_{k(t-1)} + 0.6\kappa_{kt}$ and the speech signal was estimated as $\hat{x}_{kt} = \bar{\kappa}_{kt} y_{kt}$. This smoothing slightly improves the quality of the estimated speech signal by smoothing out the gain fluctuations. For the BNMF approach [25], 60 basis vectors for speech and 100 basis vectors for babble were trained using the same training material as explained above. For this method, an informative prior was only used for babble NMF coefficients since applying informative prior for speech NMF coefficients did not result in better noise reduction performance, as also mentioned in [25].

B. Evaluations

In this section, we evaluate the proposed system and compare its performance with that of BNMF [25] and ETSI front end Wiener filtering [39]. First we present a general comparison of methods, and then some specific aspects are highlighted. Finally, the results of the subjective listening tests are given.

1) Objective Evaluation of the Noise Reduction: Five different objective measures were considered for the evaluation: (1) source to distortion ratio *(SDR)* [40] that represents the overall quality of speech; (2) long-term signal to noise ratio (*SNR*); (3) segmental SNR (*SegSNR*) [41, ch. 10], which was limited to the range [-10 dB, 30 dB]; (4) spectral distortion (*SD*) [42], for which the time-frames with powers 40 dB less than the long-term power level were excluded from the evaluations; (5) perceptual evaluation of speech quality (*PESQ*) [43]. The evaluation is performed at three input *SNR*s: 0, 5, and 10 dB.

The results are presented in Fig. 2. For *SDR*, *SNR*, and *SegSNR* the improvements in dB (e.g. $\triangle SDR = SDR_{\text{enhanced}} - SDR_{\text{noisy}}$) are shown in this figure for readability. For *PESQ* and *SD* the actual values for the enhanced signals and for the noisy input signal are shown. A high degree of consistency can be seen between the different measures. The results show that the two model-based approaches lead to much better improvements than the Wiener filtering. The proposed method outperforms the BNMF in all of the input *SNR*s in the sense of *SDR*, *SNR*, *SegSNR*, and *SD*. For *PESQ*, gamma-NHMM results to slightly better *PESQ* improvement at 0 dB input *SNR*, while BNMF leads to slightly better improvements at 5 and 10 dB *SNR*s. However, the difference between *PESQ* values for two algorithms is marginal in all three *SNR*s.

[2]By real time, we mean that the processing of the current frame finishes before the next frame arrives.

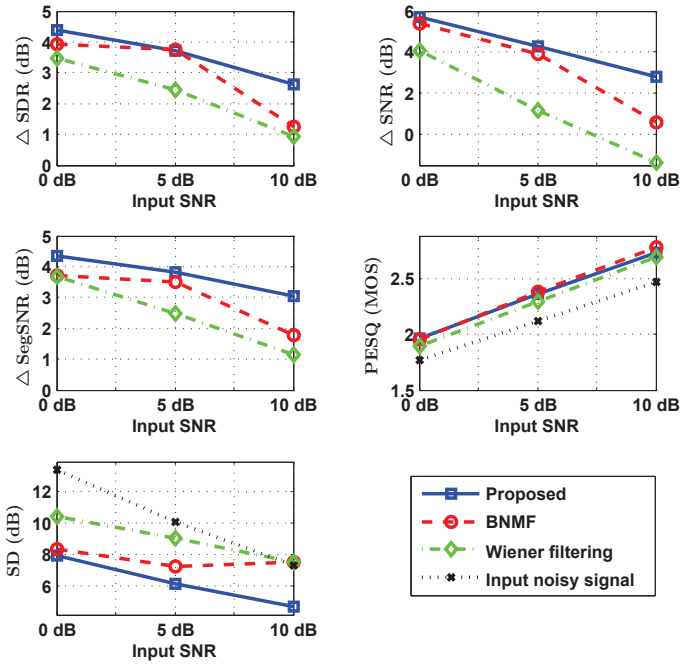


Fig. 2: Objective evaluation of the application of the proposed babble model in noise reduction. $\triangle$ is used to show the improvements gained by the noise reduction algorithms, e.g., $\triangle SDR = SDR_{\text{enhanced}} - SDR_{\text{noisy}}$.

2) Effect of Systems on Speech and Noise Separately: A desired feature of a noise reduction system is that the speech signal remains undistorted. In order to compare this aspect of the algorithms, segmental speech *SNR* ($SNR_{\text{seg-sp}}$), and segmental noise reduction (*SegNR*) [44] were measured in a shadow filtering framework. Hence, the enhancement filter was obtained using the input noisy signal (as it was done in VI-B1), and it was applied to the clean speech and noise components of the input noisy signal, separately. The output speech and noise signals were compared to the corresponding inputs to compute these two measures. For both measures a high value is desired, and $SNR_{\text{seg-sp}}$ is inversely proportional to the speech distortion.

The results are shown in Fig. 3. As it can be seen in the figure, the proposed system leads to a higher segmental speech *SNR* (less distortion) in all of the input *SNR*s. Also, the sum of the $SNR_{\text{seg-sp}}$ and *SegNR* is the highest for the proposed method.

3) Effect of the Number of Speakers in Babble: It is well known that the performance of the model-based noise reduction systems that are trained for a specific signal degrades when there is a mismatch between training and testing. Therefore, in the case of a mismatch, the standard Wiener filter might outperform the model-based approaches since it is not restricted to any specific noise type. In this part, we investigate the performance of the noise reduction algorithms as a function of the number of speakers

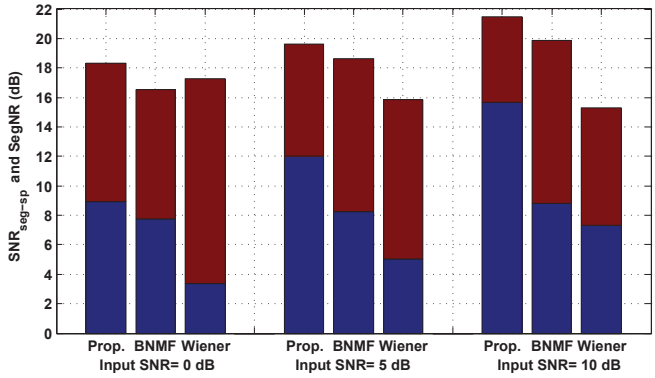


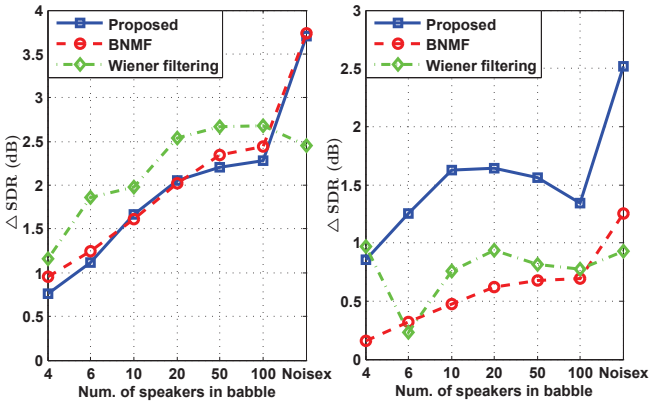
Fig. 3: Stacked presentation of the segmental speech *SNR* ($SNR_{\text{seg-sp}}$, bottom), and the segmental noise reduction (*SegNR*, top).

in the babble. For the experiments, an artificial babble was generated by adding waveforms of different speakers from the TIMIT database, with equal speech level for all of the speakers. The number of speakers in generating babble were chosen as $M \in \{4, 6, 10, 20, 50, 100\}$. Moreover, the babble noise from NOISEX-92 is also considered in the evaluation for comparison.

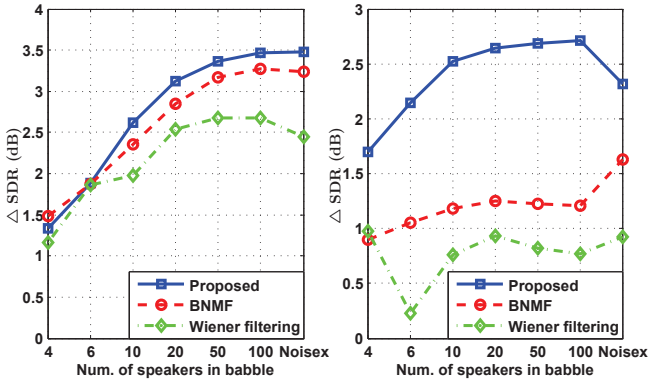
For the proposed method and BNMF, two babble models were trained using (1) only the NOISEX-92, (2) both the NOISEX-92 and 10-person babble noise (different from the test signal). Also, for the proposed system, we trained $\ddot{N} = 50$ states for the babble for both of the models since a pilot experiment indicated that $\ddot{N} = 10$ was insufficient in this case because of the high non-stationarity of the noise.

Improvements gained in the source to distortion ratio *($\triangle$SDR)* are shown in Fig. 4 for two input *SNR*s, 5 and 10 dB. Fig. 4a shows the results using the babble model that is trained on only the NOISEX-92. Looking at the 5 dB input *SNR* scenario (left-hand side of Fig. 4a), it can be seen that even though the performance of the model-based approaches is much better when exposed to the NOISEX-92 babble noise, the ETSI Wiener filter gives a better result in the other types of babble noise. Fig. 4b shows the results using the babble model that is trained using both the NOISEX-92 and 10-person artificial babble noise. Here, the performance of the model-based approaches is slightly reduced for the NOISEX-92 babble noise, but in general their performance is significantly improved (especially for the proposed method). This also implies that if the proposed method is combined with another system that estimates the number of the speakers from the observed babble signal (for example [5]), the performance might be improved further.

4) Cross-predictive Test for Model Fitting: A cross-predictive test was carried out in which both of the speech and babble models from the proposed and the BNMF frameworks were applied to the speech and



(a) Babble models trained using only NOISEX-92. The results are shown for two input *SNR*s 5 dB (left) and 10 dB (right).



(b) Babble models trained using both the NOISEX-92 and 10-person babble (different from the testing signals). The results are shown for two input *SNR*s 5 dB (left) and 10 dB (right).

Fig. 4: Performance of the noise reduction algorithms as a function of the number of speakers in the babble.

babble signals (as separate inputs) in a predictive way. Here, the goal is to investigate whether the babble (speech) model fits to the babble (speech) signal better than the speech (babble) model. Two measures were computed to compare the input and the estimated signals. To compare the signals in the spectral domain, spectral distortion (*SD*) [42] was measured. To compare the input and estimated waveforms, segmental SNR (*SegSNR*) [41, ch. 10] was measured. To reconstruct the output waveforms, the NMF representations of the spectrograms together with the phase information from the input signal were fed into the inverse DFT. For the proposed method, the gamma-NHMM representation (similar to Subsection II-B) was used to obtain the NMF approximation, and for the BNMF that was achieved by multiplying

TABLE I: A cross-predictive test for the different models. The specified signal is fed as the input signal to the given model and the quality of the reconstructed signal is measured. The results are averaged over the test set explained in Subsection VI-A.

(a) Spectral distortion (*SD*, lower value is desired) in dB.

Proposed	Speech Model	Babble Model	BNMF	Speech Model	Babble Model
Speech Sig.	3.9	6.7	Speech Sig.	6.3	9.4
Babble Sig.	3.2	2.4	Babble Sig.	2.2	2.8

(b) Segmental SNR (*SegSNR*, higher value is desired) in dB.

Proposed	Speech Model	Babble Model	BNMF	Speech Model	Babble Model
Speech Sig.	3.2	-2.1	Speech Sig.	9.1	-2.2
Babble Sig.	4	5.3	Babble Sig.	9.2	6.5

the mean values of the posterior distributions of the basis matrix and the coefficients matrix.

The results of this predictive test are shown in TABLE I in the form of confusion matrices. If a model is good then for each type of signal (each row in the table), the best result should be found in the element on the main diagonal. Both of the measures point in the same direction, and show that in the proposed framework a better score is obtained for the speech and babble signals using the speech and babble models, respectively. However, for the BNMF, the speech model gives a better score to the babble signal than the babble model itself (shown in a red color in the table). This is because the babble spectrogram can be approximated quite well by combining the speech basis vectors freely. The result of this test is another indication of the excellence of the proposed babble model, and provides an additional explanation for the achieved results in the previous subsections.

5) Subjective Evaluation of the Noise Reduction: To assess the subjective quality of the estimated speech signal, a subjective listening test was carried out. The test setup was similar to the ITU recommendation ITU-R BS.1534-1 MUSHRA [45]. Six experienced and four inexperienced listeners (ten in total) participated in the test. The subjective evaluation was performed for three input *SNR*s (0, 5, 10 dB), and for each *SNR* seven sentences from the core test set of the TIMIT database (4 males and 3 females) were presented to the listeners. In each of the 21 listening sessions, 5 signals were compared by the listeners: (1) reference clean speech signal, (2) noisy speech signal, (3,4) estimated speech signals using the gamma-NHMM and BNMF, and (5) a hidden anchor signal that was chosen to be the noisy

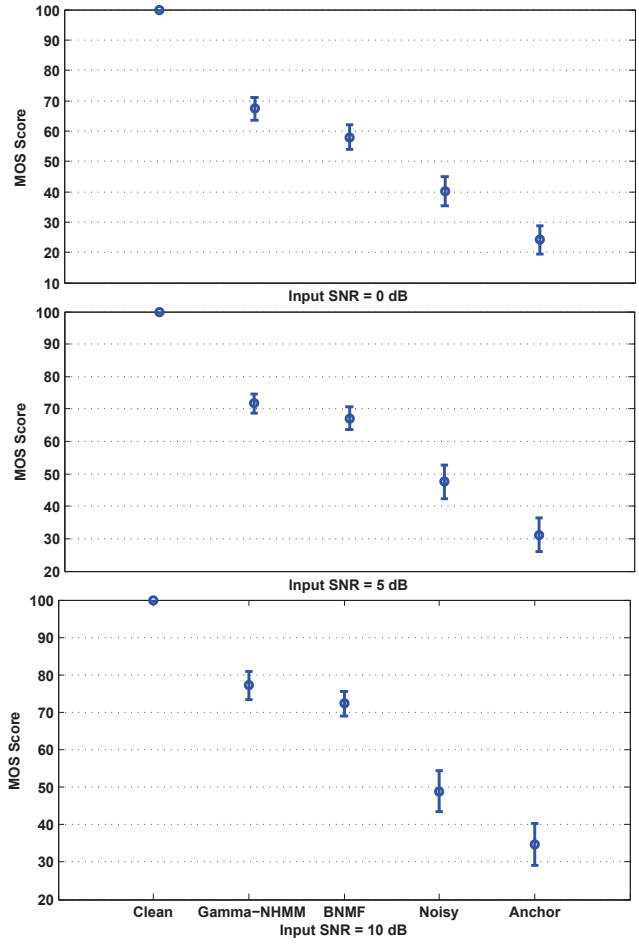


Fig. 5: Results of the MUSHRA test at 3 input *SNR*s: 0 dB, 5 dB, 10 dB (top to down) with a 95% confidence interval.

signal at a 5 dB lower *SNR* than the noisy signal processed by the systems (as suggested in [10]). The listeners were allowed to play each sentence as many times as they wanted, and they always had access to the reference signal. They were asked to rate the signals based on the global speech quality. Also, some sample signals were presented, and the graphical user interface was introduced to the listeners prior to the test procedure. The order of the signals was randomized with respect to the algorithm and input *SNR*. Each listener took around 30 minutes on average to complete the listening test.

The results of this listening test, averaged over all of the participants, with a 95% confidence interval are presented in Fig. 5. At all of the three *SNR*s the gamma-NHMM was preferred over the BNMF algorithm. For 0 dB, the difference is 9.5 MOS units, whereas for 5 dB and 10 dB, the preference is around 5 on a scale of 100. Also, both of the algorithms were preferred over the noisy input signal by at least 20 units. According to the spontaneous comments by the listeners, the remaining noise and artifacts

in the enhanced signal by the gamma-NHMM is more like a natural babble-noise while the artifacts introduced by the BNMF are more artificially modulated.

To verify the statistical significance of the preference of the gamma-NHMM algorithm, a one-tailed t-test was performed. This statistical analysis shows that the gamma-NHMM leads to a significantly better performance than the BNMF at all three *SNR*s. For 0 dB, the significance level was $p \approx 9.10^{-5}$, for 5 dB it was $p \approx 0.013$, and finally for 10 dB we obtained $p \approx 0.014$.

VII. Conclusion

As babble noise is generated by adding some different speech signals, improving the intelligibility and quality of the speech signal degraded by the babble noise has been a challenging task for a long time. In this paper, a gamma nonnegative HMM was proposed to model the normalized power spectra of babble noise in which the babble basis vectors were identical to the speech basis vectors. In the proposed models, the time-varying energy levels of speech and babble signals were modeled by gamma distributions whose scale parameters were estimated online.

The simulations show that the proposed system achieves better model recognition (i.e. babble signal gets a better score with the babble model rather than the speech model) compared to the Bayesian NMF approach. Also, the objective evaluations and the subjective MUSHRA listening test verify the excellence of the proposed noise reduction system. For instance, at 0 dB input *SNR*, the enhanced speech of the currently developed system was preferred by around 10 MOS units to the enhanced speech of the closest competing algorithm (Bayesian NMF) and by 27 to the input noisy signal in the scale of 100. Moreover, the simulations show that the proposed noise reduction scheme is less sensitive to a mismatch (varying number of speakers in babble) compared to the other competing model-based approach.

Appendix A
MAP Estimate of the Gain Variables

Problem (15) is a MAP estimator that can be solved by the standard EM algorithm. Let the hidden variables for EM be $\mathbf{Z} = \{\mathbf{X}_t, \mathbf{V}_t\}$, and $\lambda = \{g'_t, h'_t\}$ be the parameters of interest. Thus, the EM help function is written to $Q(\hat{\lambda}, \lambda) = E_{\mathbf{Z}|\mathbf{Y}_t,\lambda}(\ln f(\mathbf{y}_t, \mathbf{x}_t, \mathbf{v}_t \mid \mathbf{s}_t, \hat{\lambda}) + \ln f(\hat{\lambda}))$. The terms containing $\hat{g'_t}$ can be gathered into

$$Q_{\hat{g'_t}}\left(\hat{\lambda}, \lambda\right) = E_{\mathbf{Z}|\mathbf{Y}_t,\lambda}\left(\ln f\left(\mathbf{x}_t \mid \hat{g'_t}, \bar{S}_t = i\right) + \ln f\left(\hat{g'_t}\right)\right). \tag{44}$$

Taking the derivative of (44) and setting it to zero yields the solution:

$$\hat{g'_t} = \frac{-\left(K\theta_t - \theta_t\left(\phi - 1\right)\right) + \sqrt{\left(K\theta_t - \theta_t\left(\phi - 1\right)\right)^2 + 4\theta_t C_X}}{2}, \tag{45}$$

where K is the number of frequency bins, dimension of $\mathbf{y}_t$, and $C_X = \sum_{k=1}^{K} \frac{E\left(|X_{kt}|^2|\mathbf{Y}_t,\lambda\right)}{\alpha_k b_{ki}}$. The posterior expected value of $|X_{kt}|^2$ is calculated using (22,23). The update rule for $\hat{h_t'}$ is also given as:

$$\hat{h_t'} = \frac{-\left(K\gamma_t - \gamma_t\left(\psi-1\right)\right) + \sqrt{\left(K\gamma_t - \gamma_t\left(\psi-1\right)\right)^2 + 4\gamma_t C_V}}{2}, \tag{46}$$

where $C_V = \sum_{k=1}^{K} \frac{E\left(|V_{kt}|^2|\mathbf{Y}_t,\lambda\right)}{\beta_k[\mathbf{b}\ddot{\mathbf{s}}_t]_k}$. In very rare cases in practice, the above algorithm may get stuck at a non-maximum stationary point in which the EM algorithm has to be repeated from a different initial point to obtain a local maximum of the likelihood.

The negative Hessian matrix in (16) is defined as:

$$A_{\mathbf{s}_t}(1,1) = -\frac{\partial^2\left(\ln f\left(\mathbf{y}_t, g_t, h_t \mid \mathbf{s_t}\right)\right)}{\partial g_t \partial g_t} = \frac{(\phi-1)}{g_t^2} - \sum_{k=1}^{K} \frac{\left(g_t\alpha_k\mathbf{b}_{ki}\right)^2}{\left(g_t\alpha_k\mathbf{b}_{ki} + h_t\beta_k\left[\mathbf{b}\ddot{\mathbf{s}}_i'\right]_k\right)^2}\left(1 - \frac{2\left|y_{kt}\right|^2}{\left(g_t\alpha_k\mathbf{b}_{ki} + h_t\beta_k\left[\mathbf{b}\ddot{\mathbf{s}}_i'\right]_k\right)}\right),$$

$$A_{\mathbf{s}_t}(1,2) = A_{\mathbf{s}_t}(2,1) = -\frac{\partial^2\left(\ln f\left(\mathbf{y}_t, g_t, h_t \mid \mathbf{s_t}\right)\right)}{\partial g_t \partial h_t} = -\sum_{k=1}^{K} \frac{\left(g_t\alpha_k\mathbf{b}_{ki}\right)\left(h_t\beta_k\left[\mathbf{b}\ddot{\mathbf{s}}_i'\right]_k\right)}{\left(g_t\alpha_k\mathbf{b}_{ki} + h_t\beta_k\left[\mathbf{b}\ddot{\mathbf{s}}_i'\right]_k\right)^2}\left(1 - \frac{2\left|y_{kt}\right|^2}{\left(g_t\alpha_k\mathbf{b}_{ki} + h_t\beta_k\left[\mathbf{b}\ddot{\mathbf{s}}_i'\right]_k\right)}\right),$$

$$A_{\mathbf{s}_t}(2,2) = -\frac{\partial^2\left(\ln f\left(\mathbf{y}_t, g_t, h_t \mid \mathbf{s_t}\right)\right)}{\partial h_t \partial h_t} = \frac{(\psi-1)}{h_t^2} - \sum_{k=1}^{K} \frac{\left(h_t\beta_k\left[\mathbf{b}\ddot{\mathbf{s}}_i'\right]_k\right)^2}{\left(g_t\alpha_k\mathbf{b}_{ki} + h_t\beta_k\left[\mathbf{b}\ddot{\mathbf{s}}_i'\right]_k\right)^2}\left(1 - \frac{2\left|y_{kt}\right|^2}{\left(g_t\alpha_k\mathbf{b}_{ki} + h_t\beta_k\left[\mathbf{b}\ddot{\mathbf{s}}_i'\right]_k\right)}\right).$$

APPENDIX B
POSTERIOR DISTRIBUTION OF THE GAIN VARIABLES

The posterior distribution of the stochastic gain variable of the speech signal in Subsection V-A, given the hidden Markov state and the observation is given by:

$$f\left(g_{r,t} \mid \bar{\mathbf{o}}_{r,t}, \bar{S}_{r,t} = i, \lambda\right) = \frac{f\left(\bar{\mathbf{o}}_{r,t} \mid g_{r,t}, \bar{S}_{r,t} = i, \lambda\right) f(g_{r,t} \mid \lambda)}{f\left(\bar{\mathbf{o}}_{r,t} \mid \bar{S}_{r,t} = i, \lambda\right)},$$

where $\lambda = \{\bar{\mathbf{a}}, \mathbf{b}, \boldsymbol{\alpha}, \phi, \theta\}$ is the estimated parameters from the previous iteration of the Baum-Welch algorithm. Since the denominator is constant, using (3) and (2) we get:

$$\ln f\left(g_{r,t} \mid \bar{\mathbf{o}}_{r,t}, \bar{S}_{r,t} = i, \lambda\right) \propto$$
$$\sum_{k=1}^{K}\left(-\alpha_k \ln g_{r,t} - \frac{\bar{o}_{r,kt}}{g_{r,t} b_{k,i}}\right) +$$
$$(\phi - 1)\ln g_{r,t} - \frac{g_{r,t}}{\theta_r} =$$
$$-\frac{1}{\theta_r} g_{r,t} + \left(\phi - 1 - \sum_{k=1}^{K}\alpha_k\right)\ln g_{r,t} - \left(\sum_{k=1}^{K}\frac{\bar{o}_{r,kt}}{b_{k,i}}\right)\frac{1}{g_{r,t}}. \tag{47}$$

Eq. (47) corresponds to a generalized inverse Gaussian (GIG) distribution [46] with parameters $\vartheta = \phi - \sum_{k=1}^{K}\alpha_k$, $\rho = \frac{1}{\theta_r}$, and $\tau = \sum_{k=1}^{K}\frac{\bar{\mathbf{o}}_{r,t}}{b_{k,i}}$. The GIG distribution is generally defined as:

$$\ln \mathrm{GIG}\left(g; \vartheta, \rho, \tau\right) = -\rho g + (\vartheta - 1)\ln g - \frac{\tau}{g} +$$
$$\frac{\vartheta}{2}\ln\rho - \ln 2 - \frac{\vartheta}{2}\ln\tau - \ln \mathcal{K}_{\vartheta}\left(2\sqrt{\rho\tau}\right),$$

for $g \geq 0, \rho \geq 0$, and $\tau \geq 0$. Here, $\mathcal{K}_{\vartheta}(\cdot)$ denotes a modified Bessel function of the second kind. The required expectations are given as [46]:

$$E\left(G\right) = \frac{\mathcal{K}_{\vartheta+1}\left(2\sqrt{\rho\tau}\right)\sqrt{\tau}}{\mathcal{K}_{\vartheta}\left(2\sqrt{\rho\tau}\right)\sqrt{\rho}}, \tag{48}$$

$$E\left(G^{-1}\right) = \frac{\mathcal{K}_{\vartheta-1}\left(2\sqrt{\rho\tau}\right)\sqrt{\rho}}{\mathcal{K}_{\vartheta}\left(2\sqrt{\rho\tau}\right)\sqrt{\tau}}, \tag{49}$$

$$E\left(\ln G\right) = \frac{\frac{\partial \mathcal{K}_v\left(2\sqrt{\rho\tau}\right)}{\partial v}\big|_{v=\vartheta}}{\mathcal{K}_{\vartheta}\left(2\sqrt{\rho\tau}\right)} + \ln\sqrt{\frac{\tau}{\rho}}. \tag{50}$$

The posterior distribution of the stochastic gain variable of the noise signal can be obtained similarly.

APPENDIX C
GRADIENT AND HESSIAN FOR BABBLE STATES

The gradient of P_2 evaluated at $\widehat{\ddot{\mathbf{s}}}'_i$, which is used in the CCCP procedure in Subsection V-B, is simply given as:

$$\nabla P_2\left(\widehat{\ddot{\mathbf{s}}}'_i\right) = \left[\frac{\partial P_2\left(\widehat{\ddot{\mathbf{s}}}'_i\right)}{\partial \widehat{\ddot{s}'_{mi}}}\right] = \sum_{r,t,k}\omega_{t,r}\left(\ddot{\mathbf{s}}'_i\right)\mathbf{b}_{k\cdot}^{\top}\left(\frac{\beta_k}{\left[\mathbf{b}\widehat{\ddot{\mathbf{s}}}'_i\right]_k}\right), \tag{51}$$

where $\mathbf{b}_{k.}$ denotes the k^{th} row of the basis matrix $\mathbf{b}$, and $'\top'$ denotes the transpose. Denoting $C\left(\mathbf{x}\right) = P_1\left(\mathbf{x}\right) + \mathbf{x}^\top \nabla P_2(\widehat{\ddot{\mathbf{s}}}_i')$, the gradient and the hessian of the cost function in (34) are also given as:

$$\nabla C\left(\mathbf{x}\right) = \left[\frac{\partial C\left(\mathbf{x}\right)}{\partial x_m}\right] = \nabla P_2\left(\widehat{\ddot{\mathbf{s}}}_i'\right) - \sum_{r,t,k} \omega_{t,r}\left(\ddot{\mathbf{s}}_i'\right) \mathbf{b}_{k.}^\top \left(\frac{\ddot{o}_{r,kt}}{[\mathbf{bx}]_k^2} E_{H_{r,t}|\ddot{\mathbf{s}}_{r,t},\lambda}\left(H_{r,t}^{-1}\right)\right), \tag{52}$$

$$\nabla^2 C\left(\mathbf{x}\right) = \left[\frac{\partial^2 C\left(\mathbf{x}\right)}{\partial x_m \partial x_n}\right] = \sum_{r,t,k} \omega_{t,r}\left(\ddot{\mathbf{s}}_i'\right) \mathbf{b}_{k.}^\top \mathbf{b}_{k.} \left(\frac{2\ddot{o}_{r,kt}}{[\mathbf{bx}]_k^3} E_{H_{r,t}|\ddot{\mathbf{s}}_{r,t},\lambda}\left(H_{r,t}^{-1}\right)\right). \tag{53}$$

ACKNOWLEDGMENT

Part of this work was supported by the EU Initial Training Network AUDIS (grant 2008-214699). The authors would like to thank W. Bastiaan Kleijn for a useful discussion about the babble model.

REFERENCES

[1] E. Cherry, "Some experiments on the recognition of speech, with one and two ears," *J. of Acoustical Society of America (JASA)*, vol. 25, pp. 975–979, 1953.

[2] B. Arons, "A review of the cocktail party effect," *J. of Acoustical Society of America (JASA)*, vol. 12, pp. 35–50, 1992.

[3] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, pp. 1875–1902, 2005.

[4] S. A. Simpson and M. Cooke, "Consonant identification in N-talker babble is a nonmonotonic function of N," *J. of Acoustical Society of America (JASA)*, vol. 118, no. 5, pp. 2775–2778, 2005.

[5] N. Krishnamurthy and J. Hansen, "Babble noise: Modeling, analysis, and applications," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 7, pp. 1394–1407, sep. 2009.

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

[7] X. Shen and L. Deng, "A dynamic system approach to speech enhancement using the H_∞ filtering algorithm," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 4, pp. 391–399, jul. 1999.

[8] J. Vermaak, C. Andrieu, A. Doucet, and S. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," vol. 10, no. 3, pp. 173–185, mar. 2002.

[9] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 13, no. 5, pp. 845–856, sep. 2005.

[10] V. Grancharov and J. S. B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, pp. 764–773, 2006.

[11] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum Mean-Square Error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, pp. 1741–1752, 2007.

[12] H. Levitt, "Noise reduction in hearing aids: An overview," *J. of Rehabilitation Research and Development*, vol. 38, pp. 111–121, 2001.

[13] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 163–176, jan. 2006.

[14] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, apr. 1992.

[15] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, sep. 1998.

[16] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 882–892, mar. 2007.

[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[18] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, pp. 1–12, 2007.

[19] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, 2009.

[20] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 3, pp. 550–563, mar. 2010.

[21] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2008, pp. 4029–4032.

[22] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, may. 2011, pp. 17–20.

[23] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new approach for speech enhancement based on a constrained nonnegative matrix factorization," in *IEEE Int. Symp. on Intelligent Signal Process. and Communication Systems (ISPACS)*, 2011.

[24] ——, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoustics (WASPAA)*, 2011, pp. 45–48.

[25] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2012, pp. 4561–4564.

[26] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. of the Royal Statistical Society. Series B (Methodological)*, vol. 46, pp. 257–267, 1984. [Online]. Available: http://www.jstor.org/stable/2345509

[27] V. Krishnamurthy and J. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Trans. Signal Process.*, vol. 41, no. 8, pp. 2557–2573, aug. 1993.

[28] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, feb. 1989.

[29] O. Cappé, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models*, ser. Springer Series in Statistics. New York, Inc. Secaucus, NJ, USA: Springer, 2005.

[30] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors," *IEEE Signal Process. Letters*, 2013, to be published.

[31] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Process.*, vol. 86, no. 4, pp. 698–709, apr. 2006.

[32] C. M. Bishop, *Pattern Recognition and Machine Learning*. NewYork: Springer-Verlag, 2006.

[33] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," U.C. Berkeley, Tech. Rep. ICSI-TR-97-021, 1997.

[34] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, pp. 29–45, 1986.

[35] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, pp. 915–936, 2003.

[36] B. K. Sriperumbudur and G. R. G. Lanckriet, "On the convergence of the concave-convex procedure," in *Advances in Neural Information Process. Systems*, 2009.

[37] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[38] E. Weinstein, M. Feder, and A. Oppenheim, "Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 9, pp. 1652–1654, sep. 1990.

[39] "Speech processing, transmission and quality aspects (STQ), distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," Tech. Rep. ETSI ES 202 050 V1.1.5, 2007.

[40] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[41] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC, 2007.

[42] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis, W.B. Kleijn, K.K. Paliwal, Eds.* New York: Elsevier, 1995, ch. 12, pp. 443–466.

[43] I.-T. P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assesment of narrowband telephone networks and speech codecs," Tech. Rep., 2000.

[44] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. on Applied Signal Process.*, vol. 2005, pp. 1110–1126, 2005.

[45] *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*, ITU-R Recommendation BS.1534-1 Std., 2001-2003. [Online]. Available: http://www.itu.int

[46] T. Kawamura and K. Iwase, "Characterizations of the distributions of power inverse Gaussian and others based on the entropy maximization principle," *J. of The Japan Statistical Society*, vol. 33, no. 1, pp. 95–104, 2003.

Nasser Mohammadiha (S'11) received the M.Sc. degree in electronics engineering from Sharif University of Technology, Tehran, Iran, in 2006. He worked on digital hardware and software design until 2008.

He is currently pursuing a Ph.D. degree in telecommunications at the Sound and Image Processing laboratory, KTH Royal Institute of Technology, Stockholm, Sweden. His research interests include speech processing, mainly speech enhancement, image processing, and statistical signal modeling. He is a student member of the IEEE.

Arne Leijon (M'10) received the MS degree in engineering physics in 1971, and the Ph.D. degree in information theory in 1989, both from Chalmers University of Technology, Gothenburg, Sweden.

He has been a professor of hearing technology at the Sound and Image Processing (SIP) Laboratory at the KTH Royal Institute of Technology, Stockholm, Sweden, since 1994. His main research interest concerns applied signal processing in aids for people with hearing impairment, and methods for individual fitting of these aids, based on psychoacoustic modeling of sensory information transmission and subjective sound quality. He is a member of the IEEE.