Minimum Mean-Square Error Estimation of Mel-Frequency Cepstral Features–A Theoretically Consistent Approach

Jesper Jensen and Zheng-Hua Tan

***Abstract*—In this work, we consider the problem of feature enhancement for noise-robust automatic speech recognition (ASR). We propose a method for minimum mean-square error (MMSE) estimation of mel-frequency cepstral features, which is based on a minimum number of well-established, theoretically consistent statistical assumptions. More specifically, the method belongs to the class of methods relying on the statistical framework proposed in Ephraim and Malah's original work ("Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, 1984). The method is general in that it allows MMSE estimation of mel-frequency cepstral coefficients (MFCC's), cepstral-mean subtracted (CMS-) MFCC's, autoregressive-moving-average (ARMA)-filtered CMS-MFCC's, velocity, and acceleration coefficients. In addition, the method is easily modified to take into account other compressive non-linearities than the logarithm traditionally used for MFCC computation. In terms of MFCC estimation performance, as measured by MFCC mean-square error, the proposed method shows performance which is identical to or better than other state-of-the-art methods. In terms of ASR performance, no statistical difference could be found between the proposed method and the state-of-the-art methods. We conclude that existing state-of-the-art MFCC feature enhancement algorithms within this class of algorithms, while theoretically suboptimal or based on theoretically inconsistent assumptions, perform close to optimally in the MMSE sense.**

***Index Terms*—Robust automatic speech recognition (ASR), speech enhancement, mel-frequency cepstral coefficient (MFCC), minimum mean-square error (MMSE) estimation.**

I. Introduction

STATE-OF-THE-ART automatic speech recognition (ASR) systems typically consist of a front-end, which tries to extract relevant information - speech features - from the observed speech signals, and a back-end that matches the speech features against pre-trained statistical acoustic models. When observed speech signals resemble the speech signals used for training the acoustic models, e.g., in terms of background noise level, reverberation level, etc., then the ASR system may work well. On the other hand, when the ASR back-end is trained with noise-free speech signals, but the observed speech signals are noisy or reverberant, i.e., a mis-matched condition, then performance may decrease dramatically, e.g., [2], [3].

Several general methodologies exist for reducing the impact of environmental noise on ASR performance. These include methods, which try to reject noise and retrieve the underlying clean speech features to be presented to the ASR back-ends, e.g., [4]–[7]. They also include model adaptation methods, which adapt the back-ends to be better in line with the observed noisy features, e.g., [8], [9], methods. Other approaches use speech features that are inherently noise robust, e.g., [10]–[12]. Finally, methods exist, e.g. based on missing feature theory, which take into account the estimation uncertainty related to a given feature, e.g., [13]–[15].

In this work we consider the problem of speech feature enhancement for environment robust ASR. More specifically, given an observation of a noisy speech signal, our goal is to find minimum mean-square error (MMSE) estimates of the speech features of the underlying noise-free speech signal. Since traditional speech features, most notably mel-frequency cepstral coefficients (MFCC's), are usually computed via short-time Fourier transform (STFT) coefficients, the problem is often approached by trying to retrieve the noise-free STFT coefficients based on their observable, noisy, counterparts. For example, a popular approach is to use well-established short-time spectral speech enhancement algorithms to estimate a clean speech magnitude spectrum or periodogram based on the available noisy observation, and then simply compute resulting cepstral features by inserting these spectral estimates into expression for noise-free cepstral features. While such "plug-in" approach is simple, and may, in fact, lead to good improvements in terms of speech recognition performance, see e.g., [6], it is theoretically sub-optimal; this is so, because MMSE optimality in, e.g., the linear power domain, does not imply optimality in the cepstral domain. A more advanced approach was proposed by Stark and Paliwal [6] who assumed that the log mel-frequency energy coefficients of the clean speech signal conditioned on the noisy observation obey a Gamma distribution. Based on this assumption closed-form expressions were derived for the MMSE estimator of the MFCC vector for each frame[1]. The

Manuscript received May 31, 2014; revised August 28, 2014; accepted November 25, 2014. Date of current version January 14, 2015. Parts of this work were published in J. Jensen and Z.-H. Tan, "A Theoretically Consistent Method for Minimum Mean-Square Error Estimation of Mel-Frequency Cepstral Features," Proc. IEEE International Conference on Network Infrastructure and Digital Content (NIDC), Sep. 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yunxin Zhao.

J. Jensen is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark, and also with Oticon A/S, 2765 Smørum, Denmark (e-mail: jje@es.aau.dk; jsj@oticon.dk).

Z. -H. Tan is with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: zt@es.aau.dk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2377591

[1]In fact, as will become clear from the present work, this estimator is also the MMSE estimator for derived cepstral features, e.g., delta- and acceleration features, although this desirable property was not noticed in [6].

Gamma distribution assumption was made primarily to be able to obtain a closed-form analytical expression for the MMSE estimator, and it was shown that it performed better than most other STFT based feature enhancement methods [6]. However, a disadvantage - at least from a theoretical perspective - is that the assumption is somewhat heuristic, and cannot be proved to be consistent with the statistical assumptions made with respect to the STFT coefficients. In [5] Yu *et al.* presented an STFT based approach for feature enhancement that attempted to find MMSE estimates of clean MFCC's based on noisy MFCC's. Operating exclusively in the mel-frequency domain leads to computational savings over noise reduction methods operating in the STFT domain, because the number of mel-frequency channels is typically an order of magnitude lower than the number of STFT channels. However, the method relies on the assumption that clean mel-spectral coefficients are statistically independent from noisy mel-cepstral coefficients from different mel-frequency channels. This assumption is invalid when mel-filters overlap in frequency, which is usually the case. No performance scores were given in terms of mean-square error (MSE) estimation performance, but the method performed well in ASR experiments. In [7] Indrebo *et al.* proposed a method for MMSE estimation of MFCC's, which also operates entirely in the MFCC domain. The method assumed the noise distortions to be additive and Gaussian in the MFCC-domain, which allowed the authors to derive an estimator in closed-form. The assumption of additive and Gaussian noise in the MFCC-domain, however, is only approximately consistent with the standard assumption of the noise being additive in the time-domain.

In this paper we focus on STFT based algorithms for MMSE estimation of cepstral features in noisy environments. Specifically, we constrain our attention to the class of algorithms, which rely on the statistical model for speech and noise STFT coefficients introduced by Ephraim and Malah in [1]. This class includes a wealth of algorithms such as the short-time spectral amplitude (STSA) - MMSE algorithm [1], the log-spectral amplitude (LSA) - MMSE algorithm [16], the STFT Wiener filter [17], and the β-order - MMSE spectral amplitude estimator [18] to mention a few. In particular, in this framework, enhancement is achieved by processing noisy STFT coefficients separately for each frequency under the assumption that (detailed mathematical definitions are given below): i) target and noise STFT coefficients are uncorrelated, ii) STFT coefficients obey Gaussian distributions, and iii) STFT coefficients are statistically independent across time and frequency, given their respective spectral variances (i.e., power spectral densities (PSDs)). These statistical assumptions have made the foundation for a large range of successful speech enhancement algorithms, see, e.g., [19], [20], and the references therein. Based on these thoroughly established statistical assumptions, we propose an approach for MMSE estimation of mel-cepstral features, including MFCC's, cepstral-mean subtracted MFCC's (CMS-MFCC's), ARMA-filtered CMS-MFCC's, velocity and acceleration coefficients. The potential advantages of the proposed approach can be summarized as follows:

- given the statistical framework, it provides theoretically correct MMSE estimates of MFCC coefficients and derived features in contrast to "plug-in" algorithms.
- it does not rely on any, potentially inconsistent, assumptions (for example, the Gamma pdf assumption made in [6] is unnecessary).
- it is versatile in that it remains optimal if some of the operations leading to the derived cepstral features are skipped, e.g., if the ARMA filtering stage is omitted. Furthermore, it is straight-forward to find MMSE estimates of speech features, where the logarithmic compression used to compute traditional cepstral features [21] is replaced by physiologically more relevant compressive non-linearities, e.g., [12]. Therefore, the proposed approach could play an important role in trying out other compressive non-linearities than the ones currently known.

However, the proposed algorithm cannot be expressed in closed-form, but involves numeric solution of a one-dimensional integral. While the method is still practically useful, existing closed-form algorithms, e.g., the MMSE algorithm proposed in [6], are computationally cheaper.

The main goal of our study is to propose an algorithm, which achieves the MMSE bound for cepstral feature estimates (within the statistical framework outlined above), in the hope that the resulting MSE improvement is reflected in an improvement in ASR performance. If significant improvements can be found, then focus could be directed towards development of computationally cheaper algorithms without sacrificing performance. If, on the other hand, improvements are more modest, i.e., the performance of existing algorithms is already close to what can be achieved, then research should be directed towards other classes of algorithms.

The paper is organized as follows. In Section II, we introduce the signal model, basic assumptions, and notation. Section III reviews the theoretical expressions for mel-frequency speech features and presents a general expression for the MMSE estimator of these features. Section IV presents an algorithm for MMSE estimation of compressed mel-spectral coefficients, which serves as a basis for all estimators proposed in this paper. Section V describes implementation details, while Section VI presents simulation results with the proposed algorithms and other state-of-the-art algorithms. Finally, Section VII concludes the work.

II. Signal Model and Notation

Let us consider the following additive noise signal model

$$x(n) = s(n) + d(n),$$

where $x(n)$, $s(n)$, and $d(n)$ denote the noisy observation, the clean speech signal, and the additive noise signal, respectively, and where n is a discrete-time index.

A time-frequency domain signal representation is obtained by dividing the input signals into successive, overlapping analysis frames, applying an analysis window $h_a(n)$, and transforming the time-domain frames to the frequency domain using a Discrete Fourier Transform (DFT). The resulting STFT coefficients for the noisy speech signal are given by

$$X(k,m) = \sum_{n=0}^{N-1} x(mL_a + n) h_a(n) e^{-j2\pi kn/N}, \qquad (1)$$

where k and m denote the frequency bin index and the frame index, respectively, L_a is the frame shift in samples, and N is the DFT order. The STFT coefficients $S(k,m)$ and $D(k,m)$ for the clean and noise signal, respectively, are defined in an identical manner, so that we can write

$$X(k,m) = S(k,m) + D(k,m).$$

We consider $X(k,m)$, $S(k,m)$, and $D(k,m)$ complex-valued, zero-mean random variables, and assume that speech $S(k,m)$ and noise $D(k,m)$ STFT coefficients are uncorrelated with each other. Let $\lambda_S(k,m) = E|S(k,m)|^2$, $\lambda_W(k,m) = E|D(k,m)|^2$, and $\lambda_X(k,m) = E|X(k,m)|^2$ denote the spectral variances of the clean, noise, and noisy STFT coefficients, respectively, and observe that $\lambda_X(k,m) = \lambda_S(k,m) + \lambda_D(k,m)$. We make the standard assumptions that $S(k,m)$, $D(k,m)$, and hence $X(k,m)$ are Gaussian random variables, which are conditionally independent across time and frequency, given their respective spectral variances, e.g., [1], [22]. Finally, denote by $\xi(k,m) = \lambda_S(k,m)/\lambda_D(k,m)$ and $\zeta(k,m) = |X(k,m)|^2/\lambda_D(k,m)$ the *a priori* and *a posteriori* SNR, respectively [1], [23].

III. MMSE Estimation of Cepstral Features

In this section we derive a general expression for the MMSE estimator of any of the MFCC's, cepstral-mean subtracted MFCC's (CMS-MFCC's), ARMA-filtered MFCC's, velocities and accelerations. To do so, we first review expressions for the cepstral features in terms of clean STFT coefficients $S(k,m)$. Similar expressions hold for the noisy STFT coefficients $X(k,m)$.

The l'th mel spectral coefficient in the m th frame is defined as [9], [21]

$$M_S(l,m) = \sum_k w(k,l)|S(k,m)|^2, \quad l = 0,\ldots,L-1, \quad (2)$$

where $w(k,l) \geq 0$ is the k th coefficient of the l th triangular mel band pass filter; for later use, let Z_l denote the frequency bin index set for which $w(k,l) > 0$, i.e., frequency bins corresponding to the support of the l th triangular mel-spectrum bandpass filter. Log-mel spectral coefficients follow as

$$P_S(l,m) = g\left(M_S(l,m)\right), \quad (3)$$

where

$$g(y) = \log(y). \quad (4)$$

Alternatively, physiologically more relevant compressive non-linearities may be used, e.g., a power non-linearity of the form [12]

$$g(y) = y^\beta, \quad (5)$$

with $0 < \beta < 1$. The i th MFCC in the m th frame, $C_S(i,m)$, is given by

$$C_S(i,m) = \sum_{l=0}^{L-1} v(l,i)P_S(l,m), \quad i = 0,\ldots,I-1, \quad (6)$$

where $v(l,i)$ are coefficients of the Discrete Cosine Transform, and I is the number of MFCC's. Then, CMS-MFCC's are found by subtracting from a given cepstral coefficient, the temporal mean of that coefficient, that is

$$\bar{C}_S(i,m) = C_S(i,m) - \frac{1}{\bar{M}}\sum_{\bar{m}} C_S(i,\bar{m}), \quad (7)$$

where $\bar{M}$ is the number of cepstral coefficient in the temporal average. ARMA-filtered CMS-MFCC's are found as [24, Eq. (12)]

$$\tilde{C}_S(i,m) = c_{ARMA} \times \left(\sum_{t=-\tilde{M}_1^\gamma}^{-1} \gamma_t \tilde{C}_S(i,m+t) + \sum_{t=-\tilde{M}_1^\beta}^{\tilde{M}_2^\beta} \beta_t \bar{C}_S(i,m+t) \right). \quad (8)$$

Finally, velocity coefficients $\Delta\tilde{C}_S(i,m)$ are defined as the slope of a straight line fitted to successive $\tilde{C}_S(i,m)$'s, leading to [9]

$$\Delta\tilde{C}_S(i,m) = \frac{\sum_{t=-p}^{p} t\tilde{C}_S(i,m+t)}{\sum_{t=-p}^{p} t^2}. \quad (9)$$

In a similar manner, acceleration coefficients $\Delta^2\tilde{C}_S(i,m)$ are found as the slope of a straight line, fitted to successive $\Delta\tilde{C}_S(i,m)$ values, i.e.,

$$\Delta^2\tilde{C}_S(i,m) = \frac{\sum_{t=-p}^{p} t\Delta\tilde{C}_S(i,m+t)}{\sum_{t=-p}^{p} t^2}. \quad (10)$$

We now present a general expression for the MMSE estimator of any of these quantities. The key observation in the present context is that MFCC's, CMS-MFCC's, ARMA-filtered CMS-MFCC's, velocities, and accelerations are all linear combinations of compressed mel spectral coefficients $P_S(l,m)$. Note that this still holds, if the order of some of the operations is changed, e.g., if velocities are computed from MFCC's and not from ARMA-filtered CMS-MFCC's. Let

$$J = \sum_m \sum_l \alpha(l,m)P_S(l,m)$$

denote any such linear combination. Furthermore, let $\mathcal{X}(l,m)$ denote a vector whose entries are the set of noisy STFT coefficients that (under the statistical assumptions outlined in Section II) carry information about the specific compressed mel spectral coefficient $P_S(l,m)$, i.e., $\mathcal{X}(l,m) = \{X(k,m), k \in Z_l\}$, where the bin index set Z_l was defined after Eq. (2). Similarly, let $\mathcal{X}$ denote a vector whose entries are the total set of noisy STFT coefficients $X(k,m)$ that carry information about the total set of $P_S(l,m)$ factors in the sum. For example, for the i th MFCC $C_S(i,m)$, vector $\mathcal{X}$ consists of all noisy STFT coefficients needed to compute $C_X(i,m)$. Finally, recall that the minimum mean-square error estimate $\hat{J}$ of the linear combination J is identical to the conditional mean, e.g., [25], i.e., the ensemble average of J conditioned on all noisy observations carrying information about J. Then, the MMSE estimate may be written as

$$\begin{aligned} \hat{J} &= E\left(J|\mathcal{X}\right) \\ &= E\left(\sum_m \sum_l \alpha(l,m)P_S(l,m)|\mathcal{X}\right) \\ &= \sum_m \sum_l \alpha(l,m)\hat{P}_S(l,m) \end{aligned} \quad (11)$$

where

$$\hat{P}_S(l,m) \triangleq E\left(P_S(l,m)|X(l,m)\right), \tag{12}$$

denotes the MMSE estimate of $P_S(l,m)$.

Eq. (11) implies that in order to obtain MMSE estimates of $C_S(i,m)$, $\bar{C}_S(i,m)$, $\tilde{C}(i,m)$, $\Delta\tilde{C}_S(i,m)$, and $\Delta^2\tilde{C}_S(i,m)$, we simply need to find MMSE estimates $\hat{P}_S(l,m)$ of the compressed mel-spectral coefficients, and then form the relevant linear combinations.

IV. MMSE Estimation of Compressed Mel Spectral Coefficients

By inserting Eq. (3) in Eq. (12), the MMSE estimate $\hat{P}_S(l,m)$ of the compressed mel spectral coefficient $P_S(l,m)$ is given by

$$\hat{P}_S(l,m) = E\left\{ g\left(\sum_{k\in Z_l} w(k,l)|S(k,m)|^2 \right) |\mathcal{X}(l,m) \right\}. \tag{13}$$

Denote by $\mathcal{S}(l,m)$ a vector of all clean STFT coefficients, which contribute to $P_S(l,m)$, that is $\mathcal{S}(l,m) = \{S(k,m), k \in Z_l\}$. Furthermore, let $f(\mathcal{S}(l,m)|\mathcal{X}(l,m))$ denote the vector probability density function (pdf) of the clean STFT coefficients in vector $\mathcal{S}(l,m)$ conditioned on the noisy STFT coefficients in vector $\mathcal{X}(l,m)$. Then, Eq. (13) may be re-written as

$$\hat{P}_S(l,m) = \int_{S(l,m)} g\left(\sum_{k\in Z_l} \omega(k,l)|S(k,m)|^2 \right) \times f(\mathcal{S}(l,m)|\mathcal{X}(l,m))d\mathcal{S}(l,m), \tag{14}$$

for $l = 0,\ldots,L-1, \quad m = 0,\ldots$, where the integral is over the elements in $S(l,m)$. Unfortunately, this integral is complicated to evaluate analytically based on the statistical assumptions made so far, for any of the considered non-linearities $g(\cdot)$. Instead, we evaluate Eq. (14) numerically by drawing realizations of the vector random variable $\mathcal{S}(l,m)|\mathcal{X}(l,m)$ and approximating the integral in Eq. (14) by a sum. To this end, observe that under our distributional assumptions, the pdf $f(\mathcal{S}(l,m)|\mathcal{X}(l,m))$ is Gaussian, and is given by

$$f(\mathcal{S}(l,m)|\mathcal{X}(l,m)) = \prod_{k\in Z_l} f(S(k,m)|X(k,m)), \tag{15}$$

because STFT coefficients are conditionally independent across frequency, given their variances. Furthermore, the pdfs $f(S(k,m)|X(k,m))$ are scalar, circular symmetric, complex-valued, Gaussian, i.e.,

$$f(S(k,m)|X(k,m)) = \frac{1}{\pi\lambda_{S|X}(k,m)} \times \exp\left(-\frac{1}{\lambda_{S|X}(k,m)}|S(k,m) - \mu_{S|X}(k,m)|^2 \right) \tag{16}$$

with known mean

$$\mu_{S|X}(k,m) = \frac{\lambda_S(k,m)}{\lambda_X(k,m)}X(k,m),$$

and variance

$$\lambda_{S|X}(k,m) = \frac{\lambda_S(k,m)\lambda_D(k,m)}{\lambda_X(k,m)}.$$

So, a single realization of the vector random variable $\mathcal{S}(l,m)|\mathcal{X}(l,m)$ may simply be created by drawing realizations $\breve{S}^j(k,m)$, where the superscript j is a realization index, of independent scalar, complex random variables according to $f(S(k,m)|X(k,m))$ in Eq. (16) and stacking them in a vector. Then, the realization of the corresponding compressed mel spectral coefficient is given by

$$\breve{P}_S^j(l,m) = g\left(\sum_k w(k,l)|\breve{S}^j(k,m)|^2 \right). \tag{17}$$

Assume that N_{real} such independent realizations $\breve{P}_S^j(l,m), j = 0,\ldots,N_{real}-1$ are drawn. Then the MMSE estimate $\hat{P}_S(l,m)$ of the compressed mel spectral coefficient is approximated as

$$\hat{P}_S(l,m) \approx \frac{1}{N_{real}} \sum_{j=0}^{N_{real}-1} \breve{P}_S^j(l,m). \tag{18}$$

Note that by the law of large numbers [26], this approximation can be made arbitrarily accurate by increasing N_{real}; the variance of the estimate decreases exponentially with N_{real}, since it is an average of independently drawn random variables $\breve{P}_S^j(l,m)$, e.g., [25]. Also note that this procedure facilitates any compressive non-linearity, e.g., $g(y) = log(y)$ or $g(y) = y^\beta$ (Eqs. (4), (5)).

V. Implementation and Algorithm Outline

Analysis frames of length 200 samples (corresponding to 25 ms at a sample rate of 8 kHz) are Hamming windowed, and zero-padded, before an $N = 256$ point DFT is applied in Eq. (1)[2]. The frame shift is $L_a = 80$ samples (10 ms). The weights $w(k,l)$ in Eq. (2) implement ETSIs Aurora MFCC standard [27], where the number of filter bank channels is $L = 23$, the lowest frequency filter is centered at 125 Hz and has a bandwidth of 125 Hz, while the highest frequency filter is centered at 3657 Hz and has a bandwidth of 656 Hz.

An estimate $\hat{\lambda}_D(k,m)$ of the noise spectral variance $\lambda_D(k,m)$ is computed during a 100 ms noise-only signal region preceding speech activity (using an ideal voice activity detector (VAD)), and is assumed constant across the speech sentence. The a priori SNR $\xi(k,m)$ is estimated using the decision-directed approach [1], implemented as

$$\xi(k,m) = \max(\alpha_{dd}\frac{\widehat{A^2}(k,m-1)}{\hat{\lambda}_D(k,m)} + (1-\alpha_{dd})(\zeta(k,m)-1), \xi_{min}), \tag{19}$$

where $\widehat{A^2}(k,m)$ is the MMSE estimate of $|S(k,m)|^2$, which is given by [28]

$$\widehat{A^2}(k,m) = \left(\frac{\xi(k,m)}{1+\xi(k,m)}\right)^2 \times \left(1 + \frac{1+\xi(k,m)}{\xi(k,m)\zeta(k,m)}\right)|X(k,m)|^2. \tag{20}$$

Furthermore, $\alpha_{dd} = 0.98$, and $\xi_{min} = 0.0316$ corresponding to -15 dB.

[2]Note that $h_a(n)$ is then a 200-point Hamming window followed by 56 zeroes.

The speech spectral variance $\lambda_S(k,m)$ is estimated via the a priori SNR $\xi(k,m)$ and the estimate $\hat{\lambda}_D(k,m)$ of $\lambda_D(k,m)$ as

$$\hat{\lambda}_S(k,m) = \xi(k,m)\hat{\lambda}_D(k,m).$$

The spectral variance of the noisy signal is estimated as $\hat{\lambda}_X(k,m) = \hat{\lambda}_S(k,m) + \hat{\lambda}_D(k,m)$. The Discrete Cosine Transform coefficients $v(l,i)$ in Eq. (6) are entries of a Type 2 (orthogonal) DCT matrix, and we retain $I = 13$ cepstral coefficients. Finally, $N_{real} = 100$ realizations are used for numerical computation of the integral in Eq. (14) as a compromise between computational complexity and performance; increasing N_{real} beyond this value does not improve performance (as defined in Section VI) noteworthy, see App.A.

The proposed algorithm, which we denote as *GP-Draw* (because it is based on drawing realizations of Gaussian posterior densities), may be summarized as follows.

0) Compute estimate of noise spectral variance $\hat{\lambda}_D(k,m)$ for all m and k. If a noise tracking algorithm is used, this point is merged with step 2) below.

For each frequency index $k = 0,\ldots,N-1$, and for increasing frame indices, $m = 0,1,\ldots,$

1) Compute a priori SNR $\xi(k,m)$. For the first frame ($m = 0$), use $\xi(k,m) = |X(k,m)|^2/\hat{\lambda}_D(k,m)$. Otherwise, use Eq. (19).
2) Estimate spectral variances $\hat{\lambda}_S(k,m)$, and $\hat{\lambda}_X(k,m)$.
3) For each noisy STFT coefficient $X(k,m)$, draw N_{real} independent complex-Gaussian scalar realizations $\breve{S}^j(k,m)$ according to $f(S(k,m)|X(k,m))$, Eq. (15).
4) Compute N_{real} realizations of compressed mel spectral coefficients $\breve{P}_S^j(l,m)$, Eq. (17), for $j = 0,\ldots,N_{real}-1$, and $l = 0,\ldots,L-1$.
5) Compute MMSE estimates $\hat{P}_S(l,m)$ of compressed mel spectral coefficients $P_S(l,m)$ by averaging across realizations, Eq. (18).

Finally,

6) Compute MMSE estimates of MFCC's, CMS-MFCC's, ARMA-filtered CMS-MFCC's, velocities, and accelerations by forming the relevant linear combinations of $\hat{P}_S(l,m)$, i.e., replacing $\hat{P}_S(l,m)$ for $P_S(l,m)$ in Eqs. (6), (7), (8), (9), and (10).

VI. Results

We compare the performance of the proposed MMSE MFCC estimator with state-of-the-art methods from the literature, both in terms of estimation accuracy, and in terms of performance in automatic speech recognition experiments.

First, we consider an estimator–denoted here by *EM84*–based on Ephraim-Malah's original minimum mean-square error short-time spectral amplitude (MMSE-STSA) algorithm [1]. This algorithm produces estimates $\hat{A}_{MMSE}(k,m)$ of clean short-time *magnitude* spectra $|S(k,m)|$. The corresponding estimates of the compressed mel spectrum and MFCC's are obtained by replacing $|S(k,m)|$ by $\hat{A}_{MMSE}(k,m)$ in Eq. (2) and subsequently applying Eqs. (3) and (6)–(10). Hence, *EM84* is a "plug-in" algorithm. Secondly, we include the method proposed by Stark and Paliwal [6], which we refer to here as *SP*. We excluded the speech presence uncertainty (SPU) framework proposed there. Finally, to include a more recent spectral estimation method, we consider the method in [29] (with parameters $(\gamma,\nu) = (2, 0.45)$), which estimates the clean short-term spectral amplitudes based on a *super-Gaussian* prior, rather than the Gaussian prior underlying the statistical framework of [1]. As with *EM84*, the resulting spectral estimates are plugged into the expressions for the cepstral features, and as with *EM84*, this method is sub-optimal from a theoretical perspective. We refer to this super-Gaussian method as *SG*.

All algorithms are implemented using the decision-directed approach for *a priori* SNR estimation given by Eqs. (19), (20).

A. Performance - Estimation Accuracy in Terms of MSE

Noisy speech signals are generated artificially by adding noise signals to clean speech signals. The speech material consists of 250 speech sentences from the TIMIT data base [30] spoken by 13 female and 12 male speakers (10 sentences each). The noise signals encompass i) stationary, speech shaped noise (ssn), generated by passing white Gaussian noise through an all-pole filter fitted to the long-term spectrum of the speech signal in question, ii) car noise (car) recorded in a car cabin at 70 km/h, and iii) speech babble (bbl) from the Noisex data base [31]. All signals are downsampled to a rate of 8 kHz. The noise signal is scaled to obtain a given desired input signal-to-noise ratio (SNR) and added to the speech signal. Then, the noisy signal is pre-emphasized using a first-order IIR filter with coefficient $\alpha_s = 0.97$.

Reporting estimation performance for MFCC's and all derivatives is not practical. Instead, we concentrate on the mean-square estimation error for MFCCs only. This choice may be motivated by the fact that all derivative features are temporal linear combinations of MFCC's, and with MFCC's, which are temporally statistically independent, the MSE of any derivative feature is simply a linear combination of the MFCC-MSE. Hence, we measure the estimation MSE for each MFCC index via the following normalized MSE,

$$\varepsilon(i) = \frac{\frac{1}{M}\sum_{m=0}^{M-1}\left(\hat{C}_S(i,m) - C_S(i,m)\right)^2}{\frac{1}{M}\sum_{m=0}^{M-1}\left(C_S(i,m)\right)^2}, \qquad (21)$$

where M denotes the total number of frames, and $\hat{C}_S(i,m)$ is the estimated MFCC. In order to condense the performance of a given estimator into a single number, we use the normalized mean square error $\varepsilon(i)$ defined above, averaged across cepstral dimensions,

$$\varepsilon = \frac{1}{I}\sum_{i=0}^{I-1}\varepsilon(i). \qquad (22)$$

Fig. 1 shows performance in terms of $\varepsilon(i), i = 0,\ldots,I-1$ for speech-shaped noise, car noise, and babble noise, respectively, for an input SNR of 10 dB. Generally speaking, *SP* and the proposed method, *GP-Draw*, show almost identical performance, *EM84* performs slightly worse, while *SG* performs better for lower cepstral indices but worse for higher cepstral indices. For speech-shaped and car noise, all methods lead to improvements for all cepstral indices. For babble noise, performance is generally much worse. This can be attributed to the fact that this noise source is somewhat non-stationary, which is in contradiction with the algorithm implementations used here. Straighforward extension of the methods with adaptive noise power spectral density tracking methods, e.g., [32], [33], is expected to improve performance for all methods in this situation.

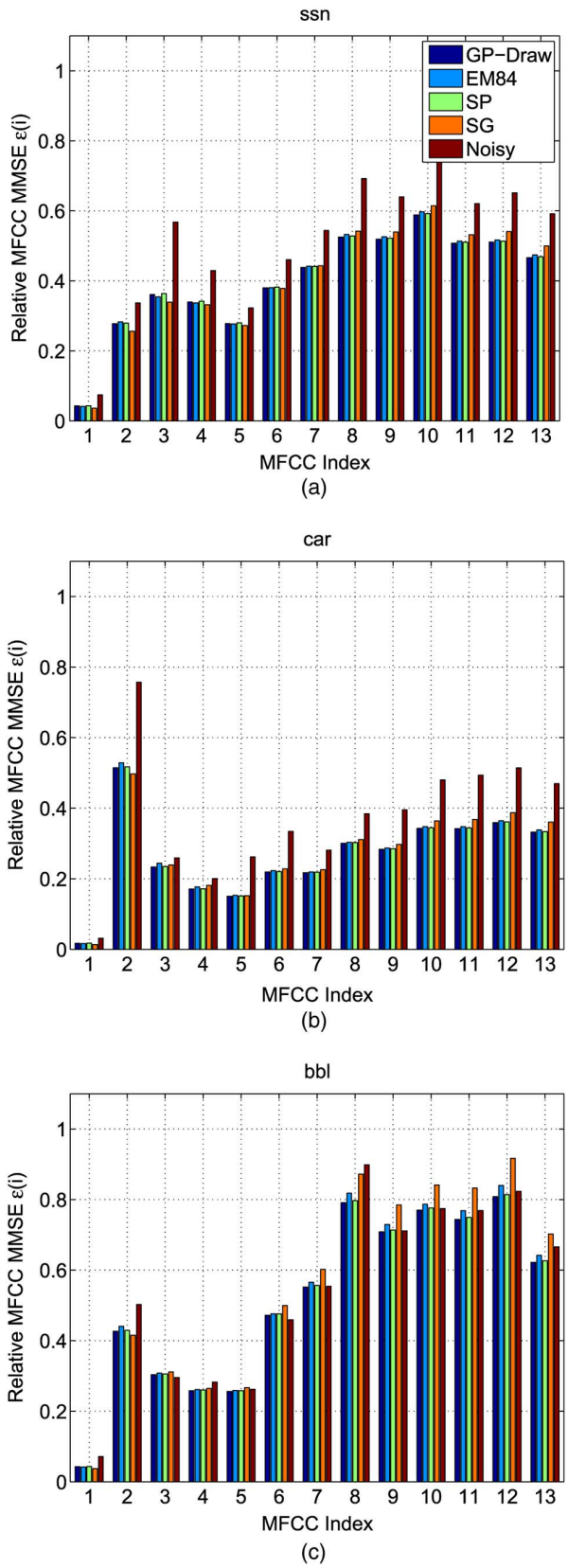

Fig. 1. Normalized mean-square estimation errors $\varepsilon(i)$ for each MFCC for input SNR of 10 dB. (a) speech-shaped noise. (b) car noise. (c) babble noise.

Fig. 2 shows estimation performance in terms of ε as a function of input SNR for the three noise sources. This figure supports the conclusion that the proposed method and *SP* manage to improve MFCC quality across input SNR for all noise sources; still, these two methods deliver almost identical performance. The *EM84* method performs slightly worse; it leads to improvements for speech-shaped and car noise, but degrades performance for low input SNRs in babble noise. Finally, *SG* performs worse in terms of MSE than the other methods.

B. Performance - Automatic Speech Recognition

In this section we use the MFCC enhancement algorithms from the previous section as feature enhancement algorithms for an automatic speech recognition system.

Speech Recognition Database and Experimental Setups: Experiments were conducted with the Aurora 2 database [34], which is the TI connected digits database artificially distorted by adding noise and using a simulated channel distortion. The sampling rate is 8 kHz. Whole word models were created for all digits using the HTK recognizer [35] and trained on clean speech data. For testing, all three test sets were used, each including clean speech and noisy speech corrupted by different types of noise with SNR values ranging from 0 to 20 dB with 5 dB intervals. The four noise types in Test Set A are subway, babble, car, and exhibition while the four types of noise in Test Set B are restaurant, station, airport, and street. Test Set C includes two noise types, subway and street, in addition to convolutional noise. Each noise type and condition has 1001 test utterances, leading to 60060 utterances in total for testing. The speech features are 12 ARMA-filtered CMS-MFCC coefficients, logarithmic energy as well as their corresponding velocity and acceleration components. To compute the ARMA-filtering in Eq. (8), we used [24, Eq. (12)] $\tilde{M}_1^{\gamma} = 2$, $\tilde{M}_1^{\beta} = 0$, $\tilde{M}_2^{\beta} = 2$, $\gamma_t = \beta_t = 1$, and $c_{ARMA} = \frac{1}{5}$, and to compute velocity and acceleration coefficients, we used $p = 2$ in Eqs. (9) and (10).

Table I summarizes the average word accuracy (WA) for Test Set A, obtained with the studied methods. All feature enhancement methods succeed in improving average performance over the noisy condition. Performance is almost identical for *SP*, *EM84*, and *GP-Draw*, while *SG* performs better. According to [6], [36] a 1.95% absolute difference in WA for the Aurora2 database is required to meet the statistical significance test ($p = 0.05$). Table I shows that *SG* meets this requirement (for the Set average) in comparison to *SP* and *GP-Draw*.

Table II shows the average WA results for Test Set A, obtained with the studied methods when a power non-linearity is used instead of the traditional log non-linearity. Using a power non-linearity increases performance quite significantly: the absolute improvements (average across the test set) for *EM84*, *GP-Draw*, and *SG* are 5.41%, 5.99%, and 3.60%, respectively, and the three methods show essentially identical performance. Performance for the noisy, unprocessed signal with logarithmic non-linearity (a WA of 69.29% as shown in Table I) is improved by more than 12%. Note that all experimental results reported in this paper are based on ARMA-filtered CMS-MFCC. As a reference, the basic MFCC with logarithmic non-linearity and without applying ARMA-filter and CMS gives a WA of 60.92% averaged across SNRs of 0 to 20 dB and across all noise types in Test Set A.

Tables III and IV show the results for Test Set B, obtained with the studied methods when the logarithmic non-linearity and power non-linearity are used, respectively. The differences

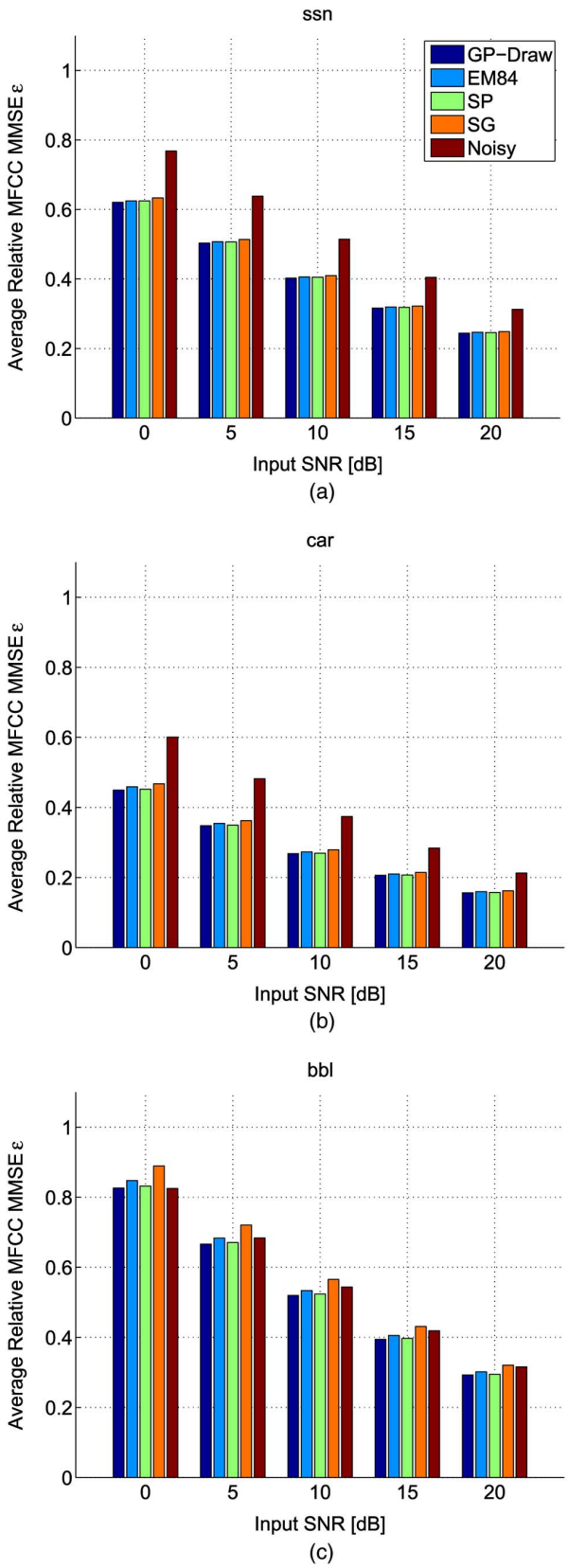

Fig. 2. Normalized mean-square estimation errors averaged across MFCC dimensions, ε as a function of input SNR. (a) speech-shaped noise. (b) car noise. (c) babble noise.

between the enhancement methods are not significant, but the improvements over the noisy, unprocessed signal are significant

TABLE I
AURORA-2A WORD ACCURACY [%]. FEATURES ARE ARMA-FILTERED CMS-MFCC'S COMPUTED WITH LOGARITHMIC NON-LINEARITY, I.E., $g(y) = \log(y)$ (EQ. (4)).

	SNR [dB]						
	0	5	10	15	20	Avg.	Clean
Subway							
Noisy	25.45	53.21	80.41	92.69	96.59	69.67	99.02
SP	36.69	68.19	86.21	94.17	96.59	76.37	98.89
EM84	38.41	69.57	86.86	94.32	96.62	77.16	98.83
GP-Draw	36.44	68.19	86.12	94.20	96.56	76.30	98.89
SG	43.48	73.04	88.03	94.47	96.90	79.18	98.83
Babble							
Noisy	27.39	60.28	84.92	94.68	97.76	73.01	98.91
SP	35.85	68.83	87.30	94.89	97.25	76.82	98.94
EM84	36.91	69.56	87.18	94.86	97.16	77.13	99.00
GP-Draw	35.70	68.53	87.27	95.01	97.28	76.76	98.94
SG	39.30	69.80	87.27	94.38	97.16	77.58	98.88
Car							
Noisy	22.79	47.69	81.24	93.56	97.64	68.58	98.96
SP	33.31	70.18	90.01	96.69	98.09	77.66	98.78
EM84	35.43	72.14	90.67	96.72	98.21	78.63	98.81
GP-Draw	33.37	70.24	89.89	96.69	98.09	77.66	98.78
SG	42.02	77.18	92.28	97.20	98.24	81.38	98.87
Exhibition							
Noisy	20.86	46.59	75.87	90.44	95.77	65.91	98.95
SP	31.04	61.65	82.84	92.41	96.02	72.79	99.01
EM84	32.06	62.85	83.09	92.50	96.02	73.30	99.04
GP-Draw	30.98	61.68	82.66	92.35	95.99	72.73	99.01
SG	35.33	64.15	84.26	92.84	95.99	74.51	99.11
Set Avg.							
Noisy	24.12	51.94	80.61	92.84	96.94	69.29	98.96
SP	34.22	67.21	86.59	94.54	96.99	75.91	98.91
EM84	35.70	68.53	86.95	94.60	97.00	76.56	98.92
GP-Draw	34.12	67.16	86.49	94.56	96.98	75.86	98.91
SG	40.03	71.04	89.96	94.72	97.07	78.17	98.92

for the log non-linearity. The absolute improvements by using a power non-linearity over a logarithmic non-linearity for *EM84*, and *GP-Draw*, are 1.48% and 2.09%, respectively, while performance for *SG* decreases by 0.27%. The absolute improvements by using a power non-linearity over a logarithmic non-linearity for *EM84* and *GP-Draw* are 1.92% and 2.72%, respectively, with the latter being significant. The relatively smaller improvement as compared with Test Set A is due to the low performance on Restaurant Noise in Test Set B. The differences for clean speech for all methods are very minor.

Table V and VI show the results for Test Set C, obtained with the studied methods when a logarithmic and power non-linearity are used, respectively. With the log non-linearity, *SG* performs significantly better than *SP* and *GP-Draw*, while with the power non-linearity, the difference between the enhancement methods is insignificant. The improvement over the noisy, unprocessed signal is significant. The absolute improvements by using a power non-linearity over a logarithmic non-linearity for *EM84*, *GP-Draw*, and *SG* are 6.94%, 7.01%, and 6.07% respectively, with all being significant. The differences for clean speech for all methods are very minor.

C. Discussion

Considering MFCC estimation performance in terms of MSE, there are only small differences between the studied methods: *SP* and *GP-Draw* show essentially identical performance, *EM84* is slightly worse, and *SG* generally shows

TABLE II
AURORA-2A WORD ACCURACY [%]. FEATURES ARE ARMA-FILTERED CMS-MFCC'S COMPUTED WITH POWER NON-LINEARITY, I.E., $g(y) = y^{\beta}, \beta = 1/15$ (EQ. (5)). *SP* ASSUMES A LOGARITHMIC NON-LINEARITY, SO NO *SP* SCORE CAN BE COMPUTED.

	SNR [dB]						
	0	5	10	15	20	Avg.	Clean
Subway							
Noisy	49.71	74.88	89.01	95.43	97.79	81.36	99.20
SP	-	-	-	-	-	-	-
EM84	59.53	82.10	91.68	95.92	97.76	85.40	99.23
GP-Draw	58.67	81.42	91.74	95.86	97.61	85.06	99.20
SG	60.45	82.90	92.23	96.19	97.82	85.92	99.17
Babble							
Noisy	33.28	64.69	86.12	95.04	98.04	75.43	99.06
SP	-	-	-	-	-	-	-
EM84	39.15	67.29	86.19	94.26	97.16	76.81	98.94
GP-Draw	39.81	68.29	86.85	94.59	97.22	77.35	98.91
SG	35.85	63.88	83.34	92.53	96.70	74.46	98.97
Car							
Noisy	41.87	74.20	90.13	96.00	98.33	80.11	99.05
SP	-	-	-	-	-	-	-
EM84	62.75	85.45	94.45	97.38	98.60	87.73	98.93
GP-Draw	61.50	84.82	93.98	97.35	98.48	87.23	98.84
SG	66.60	87.27	95.20	97.49	98.54	89.02	98.93
Exhibition							
Noisy	37.23	66.37	84.54	93.09	96.08	75.47	99.32
SP	-	-	-	-	-	-	-
EM84	44.86	69.89	85.56	93.49	95.96	77.95	99.14
GP-Draw	44.12	69.64	85.34	93.71	96.08	77.78	99.14
SG	44.74	68.81	85.44	93.52	95.80	77.66	99.23
Set Avg.							
Noisy	40.53	70.04	87.45	94.89	97.56	78.09	99.16
SP	-	-	-	-	-	-	-
EM84	51.57	76.18	89.47	95.26	97.37	81.97	99.06
GP-Draw	51.03	76.04	89.48	95.38	97.35	81.85	99.02
SG	51.91	75.72	89.05	94.93	97.27	81.77	99.08

TABLE III
AURORA-2B WORD ACCURACY [%]. FEATURES ARE ARMA-FILTERED CMS-MFCC'S COMPUTED WITH LOGARITHMIC NON-LINEARITY, I.E., $g(y) = \log(y)$ (EQ. (4)).

	SNR [dB]						
	0	5	10	15	20	Avg.	Clean
Restaurant							
Noisy	31.69	63.80	85.48	94.69	97.36	74.60	99.02
SP	36.30	65.40	84.46	92.57	96.81	75.17	98.89
EM84	37.58	65.98	84.83	92.75	96.62	75.55	98.83
GP-Draw	36.54	65.43	84.53	92.57	96.81	75.18	98.86
SG	38.59	65.77	84.13	92.11	95.89	75.30	98.83
Street							
Noisy	27.36	57.50	83.28	94.17	97.13	71.89	98.91
SP	34.76	67.56	87.36	94.62	96.89	76.24	98.94
EM84	36.00	68.41	87.48	95.71	96.86	76.69	99.00
GP-Draw	34.92	67.59	87.33	94.65	96.83	76.26	98.94
SG	40.84	70.80	88.21	94.77	96.77	78.28	98.88
Airport							
Noisy	32.06	62.81	86.70	94.87	97.29	74.75	98.96
SP	40.56	70.30	87.53	95.20	96.78	78.07	98.78
EM84	41.87	70.83	87.56	95.05	96.87	78.42	98.81
GP-Draw	40.56	69.97	87.41	95.26	96.84	78.01	98.81
SG	44.71	72.17	87.68	94.69	96.63	79.18	98.87
Train							
Noisy	26.54	56.43	83.83	94.17	97.75	71.74	98.95
SP	35.58	70.93	88.92	95.12	97.38	77.59	99.01
EM84	37.30	72.11	89.48	95.09	97.22	78.24	99.04
GP-Draw	35.58	70.75	89.02	95.12	97.38	77.57	99.01
SG	42.09	75.47	90.47	95.22	97.19	80.01	99.11
Set Avg.							
Noisy	29.41	60.14	84.82	94.48	97.38	73.24	98.96
SP	36.88	68.55	87.07	94.38	96.97	76.77	98.91
EM84	38.17	69.33	87.34	94.40	96.89	77.23	98.92
GP-Draw	36.90	68.44	87.07	94.40	96.97	76.75	98.91
SG	41.56	71.05	87.62	94.20	96.62	78.21	98.92

poorest MMSE performance. This rather small difference is somewhat surprising: the proposed method, *GP-Draw*, is an MMSE estimator based on a minimum number of assumptions, which are well-established in the area of single-channel speech enhancement. For that reason, we expect the method performs well. The *SP* method relies on an additional assumption (the Gamma pdf assumption, see above), and if this assumption is valid, *SP* is MMSE optimal as well; it is not completely surprising that this estimator works well (similar results were reported in [6]). It is, however, more surprising that *EM84* performs almost as well in terms of MSE; *EM84* is a relatively simple *ad hoc* method, which cannot claim optimality in any sense. It may be concluded that estimation accuracy in terms of MSE is not very sensitive to accurate modeling of the conditional log-mel spectral coefficients (a property offered by *SP* and *GP-Draw*). Finally, the relatively poor MMSE performance by *SG* emphasizes that good MMSE performance in the linear amplitude domain [29] does not necessarily lead to good MMSE performance in the MFCC domain.

Turning to ASR performance, feature enhancement generally improves performance. Again, there are only small performance differences between *SP*, *GP-Draw* and *EM84*, and the differences are not statistically significant ($p = 0.05$). Best performance is achieved using a power non-linearity: here all methods, including *SG* shows similar average ASR performance.

As for MSE performance, the good ASR performance for *EM84* is somewhat unexpected. One possible explanation is that the MFCC MSE performance measure does not completely correlate with ASR performance, a hypothesis, which is supported by the MMSE and ASR performance of the *SG* estimator: in other words, optimality in terms of MFCC MSE may not imply optimality in terms of ASR performance.

It is interesting to note that *SP* and *GP-Draw* perform almost identically both in terms of MSE and ASR performance. This implies that the Gamma assumption made in [6] is not only better than alternatives such as Gaussian, Log-Normal and Chi-Square assumptions [6], but is indeed close to optimal. In situations where a logarithmic non-linearity is used for computing cepstral features, *SP* is therefore an equally good and computationally cheaper alternative to *GP-Draw*.

Finally, it is clear that the ASR performance of the algorithms under study is significantly worse than the performance of state-of-the-art ASR systems, such as the ETSI Advanced Front-End (AFE), which achieves average WAs for Sets A, B, C of 87.74%, 87.19%, and 85.44%, respectively [37]. The difference may partly be explained by the fact that the ETSI AFE system is optimized for ASR performance, whereas the algorithms under study aim at MMSE optimality (in the hope that this leads to improved ASR performance). Another possible explanation is that the ETSI AFE exploits across-frequency information, which the algorithms under study in this paper do not.

TABLE IV
AURORA-2B WORD ACCURACY [%]. FEATURES ARE ARMA-FILTERED CMS-MFCC'S COMPUTED WITH POWER NON-LINEARITY, I.E., $g(y) = y^{\beta}, \beta = 1/15$ (EQ. (5)). *SP* ASSUMES A LOGARITHMIC NON-LINEARITY, SO NO *SP* SCORE CAN BE COMPUTED.

	SNR [dB]						
	0	5	10	15	20	Avg.	Clean
Restaurant							
Noisy	33.90	62.33	82.50	92.17	96.78	73.54	99.20
SP	-	-	-	-	-	-	-
EM84	34.45	58.61	78.72	88.98	95.33	71.22	99.23
GP-Draw	34.69	59.32	79.49	89.62	95.89	71.80	99.20
SG	31.26	54.53	75.31	86.25	94.04	68.28	99.17
Street							
Noisy	44.50	72.88	89.15	95.83	98.22	80.12	99.06
SP	-	-	-	-	-	-	-
EM84	51.09	76.48	90.11	96.16	97.31	82.23	98.94
GP-Draw	51.15	76.24	89.90	96.10	97.28	82.13	98.94
SG	52.33	72.12	90.08	95.80	97.28	82.52	98.97
Airport							
Noisy	41.75	69.88	87.32	94.84	97.11	78.18	99.05
SP	-	-	-	-	-	-	-
EM84	47.99	71.64	86.28	93.47	96.27	79.13	98.93
GP-Draw	47.99	71.79	86.73	93.77	96.39	79.33	98.84
SG	47.00	69.67	85.18	92.66	95.85	78.07	98.93
Train							
Noisy	38.51	69.33	87.47	94.35	97.50	77.43	99.32
SP	-	-	-	-	-	-	-
EM84	52.61	77.41	89.94	94.60	96.82	82.28	99.14
GP-Draw	51.99	77.41	89.63	94.66	96.79	82.10	99.14
SG	55.85	78.12	89.94	94.11	96.51	82.91	99.23
Set Avg.							
Noisy	39.67	68.61	86.61	94.30	97.40	77.32	99.16
SP	-	-	-	-	-	-	-
EM84	46.54	71.04	86.26	93.30	96.43	78.71	99.06
GP-Draw	46.46	71.19	86.44	93.54	96.59	78.84	99.03
SG	46.61	69.86	85.13	92.21	95.92	77.94	99.08

TABLE V
AURORA-2C WORD ACCURACY [%]. SIGNALS ARE MIRS-FILTERED. FEATURES ARE ARMA-FILTERED CMS-MFCC'S COMPUTED WITH LOGARITHMIC NON-LINEARITY, I.E., $g(y) = y^{\beta}, \beta = 1/15$ (EQ. (4)).

	SNR [dB]						
	0	5	10	15	20	Avg.	Clean
Subway							
Noisy	23.67	47.65	76.73	91.31	96.44	67.16	99.08
SP	32.24	63.13	85.29	94.11	96.50	74.25	98.93
EM84	33.10	64.91	85.91	94.32	96.75	75.00	98.96
GP-Draw	32.15	63.22	85.29	94.07	96.59	74.26	98.96
SG	37.06	69.63	87.53	94.57	96.75	77.11	99.05
Street							
Noisy	25.30	52.39	79.84	93.26	97.07	69.57	99.09
SP	34.25	63.42	84.40	94.41	96.40	74.58	98.97
EM84	35.34	64.63	85.10	94.26	96.49	75.16	99.03
GP-Draw	34.40	63.51	84.46	94.32	96.40	74.62	99.00
SG	38.27	67.53	85.85	94.20	96.58	76.49	99.00
Set Avg.							
Noisy	24.49	50.02	78.29	92.29	96.76	68.37	99.09
SP	33.25	63.28	84.85	94.26	96.45	74.42	98.95
EM84	34.22	64.77	85.51	94.29	96.62	75.08	99.00
GP-Draw	33.28	63.37	84.88	94.20	96.50	74.44	98.98
SG	37.67	68.58	86.69	94.39	96.67	76.79	99.03

VII. CONCLUSION

We presented a method for MMSE MFCC feature estimation, which is based on a minimum number of well-proven assumptions, and, which is theoretically consistent. Specifically, assuming that STFT coefficients are processed independently for each frequency, and that i) target and noise STFT coefficients are uncorrelated, ii) STFT coefficients obey Gaussian distributions, and iii) STFT coefficients are statistically independent across time and frequency, given their respective PSDs, the proposed method provides MMSE estimates of MFCC's, cepstral mean-subtracted MFCC's (CMS-MFCC's), ARMA-filtered CMS-MFCC's, velocity and acceleration coefficients. Furthermore, the proposed method is operational for other compressive non-linearities than the $\log(\cdot)$ traditionally used for MFCC computation, e.g., a power non-linearity.

TABLE VI
AURORA-2C WORD ACCURACY [%]. SIGNALS ARE MIRS FILTERED. FEATURES ARE ARMA-FILTERED CMS-MFCC'S COMPUTED WITH POWER NON-LINEARITY, I.E., $g(y) = y^{\beta}, \beta = 1/15$ (EQ. (5)). SP ASSUMES A LOGARITHMIC NON-LINEARITY, SO NO SP SCORE CAN BE COMPUTED.

	SNR [dB]						
	0	5	10	15	20	Avg.	Clean
Subway							
Noisy	39.98	68.13	85.57	93.86	96.87	76.88	99.32
SP	-	-	-	-	-	-	-
EM84	53.21	78.75	90.54	95.86	97.61	83.19	99.26
GP-Draw	51.86	77.22	89.96	95.58	97.51	82.43	99.23
SG	56.71	80.75	91.53	96.10	97.70	84.56	99.26
Street							
Noisy	39.57	68.95	85.46	94.04	97.67	77.14	99.12
SP	-	-	-	-	-	-	-
EM84	48.40	75.00	88.36	95.10	97.40	80.85	99.09
GP-Draw	47.46	74.27	88.09	95.01	97.49	80.46	99.09
SG	49.46	75.42	88.88	94.71	97.37	81.17	99.12
Set Avg.							
Noisy	39.78	68.54	85.52	93.95	97.27	77.01	99.22
SP	-	-	-	-	-	-	-
EM84	50.81	76.88	89.45	95.48	97.51	82.02	99.18
GP-Draw	49.66	75.75	89.03	95.30	97.50	81.45	99.16
SG	53.09	78.09	90.21	95.41	97.54	82.86	

In simulation experiments with speech signals contaminated by various additive noise sources, the proposed method succeeds in reducing MFCC MSE, compared to the original noisy MFCC's. In comparison with other methods based on short-term spectral coefficient estimation, it shows lower MSE than a method based on the Ephraim-Malah short-time spectral amplitude MMSE estimator [1], and a more recent method based on a *super-Gaussian* short-time spectral amplitude prior [29]. Furthermore, the proposed method was compared to the method (*SP*) by Stark and Paliwal [6], which relies on the additional assumption that the log mel-frequency energy of the clean signal conditioned on the noisy observation obeys a Gamma distribution. This method leads to essentially identical performance in terms of estimation MSE as the proposed method. The advantage of the proposed method in this situation is that it remains optimal for other non-linearities than the $\log(\cdot)$, which *SP* is restricted to, although at the cost of higher computational complexity.

In ASR experiments, all feature enhancement methods succeed in improving performance over the unprocessed baseline. Somewhat surprisingly, only small performance differences are observed between the methods under study. In fact, the *EM84* method, which is theoretically sub-optimal, performs slightly better than *SP* and *GP-Draw*, which are theoretically easier to justify (although this performance difference is not statistically significant).

The main goal of our study was to propose an STFT based algorithm for cepstral feature estimation, which is optimal in MMSE sense (given the well-proven statistical framework outlined above), in the hope that the resulting MSE improvement is reflected in an ASR improvement. The fact that MSE performance of existing schemes is quite close to that of the proposed scheme suggests that existing schemes within this class of algorithms are already almost optimal. Further improvements, however, may be achieved by a) refining the existing assumptions so that the signal model reflects the observed signals better, or b) extending the set of assumptions to increase the amount of *a priori* knowledge built into the algorithm.

Considering first a refinement of the existing assumptions, it is well-known that STFT coefficient estimation in a speech enhancement context may be improved by replacing the Gaussian STFT assumption with a super-Gaussian assumption [38]. The estimator *SG* included in the study belongs to this class, although here it was used as a "plug-in" mfcc estimator, and can therefore not claim MMSE optimality. Furthermore, since spectral amplitude estimators based on super-Gaussian priors lead to relatively modest improvements in terms of speech enhancement performance [29], [38], it may be expected that the improvements in terms of MFCC estimation performance would remain modest. Finally, to simplify the interpretation of our results, the methods considered in this paper relied on a stationary noise assumption. For non-stationary noise sources, it is expected that performance can be improved via a straightforward introduction of methods for noise power spectral tracking, e.g., [32], [33], [39].

Secondly, and perhaps more importantly, more *a priori* knowledge can be introduced in the enhancement process. For example, it may be noted that the class of STFT estimation based methods considered in this paper model STFT coefficients as conditionally independent, given speech spectral variances, which are estimated independently for each frequency band. Specifically, all methods considered here estimate the speech spectral variances using an unbiased decision-directed approach, (Eqs. (19)–(20)), applied independently to each frequency subband. In this way, however, the spectro-temporal structure of speech (and noise) spectral variances are not fully exploited, and significant performance improvements may be found by applying more advanced estimation methods, which to a larger extent make use of prior speech and noise signal knowledge. Examples of such methods include methods that exploit speech power spectral density structures across frequency, e.g., via spectral codebooks [40], via Gaussian Mixture Models (GMMs), e.g., [41], or via cepstrum smoothing techniques, e.g., [42], or methods, which exploit spectro-temporal speech (and noise) psd structures, e.g., [43]–[45], see also [46, Chap. 2] and the references therein.

Finally, in this paper we have focused on MMSE estimators of MFCC features. The MMSE criterion was partly chosen because of mathematical tractability, and partly because of lack of obvious alternatives. Our results, however, indicate that optimality in terms of MFCC MMSE does not necessarily imply optimal ASR performance. Mathematically tractable alternatives to the MMSE criterion for ASR performance prediction are important topics for future research.

APPENDIX
APPENDIX PERFORMANCE VERSUS COMPLEXITY

It is difficult to determine an appropriate value of N_{real} analytically. Instead we determine it via simulation experiments by computing estimation performance in terms of the mean-square MFCC estimation error as a function of N_{real}. For convenience, let us repeat the definition from the main text of the normalized MSE for the i th MFCC, when the MMSE-MFCC estimate, $\hat{C}_S(i,m;N_{real})$, is computed using a particular N_{real},

$$\varepsilon(i;N_{real}) = \frac{\sum_{m=0}^{M-1}\left(\hat{C}_S(i,m;N_{real}) - C_S(i,m)\right)^2}{\sum_{m=0}^{M-1} C_S(i,m)^2}. \quad (21)$$

We evaluated Eq. (21) for noisy speech constructed by adding speech shaped Gaussian noise to 100 arbitrarily selected speech signals from the TIMIT data base [30], at an SNR of 0 dB. Each speech signal was repeated with 75 independently drawn noise realizations.

Since we are mainly interested in the convergence behavior of $\varepsilon(i;N_{real})$, we plot in Fig. 3 a normalized version of $\varepsilon(i;N_{real})$,

$$\varepsilon'(i;N_{real}) = \frac{\varepsilon(i;N_{real})}{\varepsilon(i;N_{real=1})}, \quad i=0,\ldots,I-1, \quad (23)$$

with $I = 13$ MFCC's. Convergence appears to be reached with $N_{real} > 80$, although the performance loss in using lower values of N_{real} appears small. It is interesting to note that the curves in Fig. 3 arrange themselves from top to bottom as $\varepsilon'(i;N_{real}), i = 0,1,\ldots,I-1$. This can be explained if we consider the impact on the estimate $\hat{C}_S(i,m;N_{real})$ of increasing N_{real}: recall that $\hat{C}_S(i,m;N_{real})$ is constructed as a linear combination of compressed mel-spectral estimates $\hat{P}_S(l,m;N_{real})$. For low values of N_{real}, the estimate $\hat{P}(i,m;N_{real})$ has a relatively large variance; it can be considered subject to 'jitter' or noise. It appears reasonable that this jitter does not affect large-scale features of the compressed mel-spectrum; for example, the spectral envelope may be largely unchanged due to the jitter. This is in line with Fig. 3, which shows that low-index MFCC's, which primarily encode large-scale spectral features, are insensitive to N_{real}. Finer spectral details, which are encoded in higher-index MFCC's, however, may be more sensitive to the jitter, which is supported by Fig. 3. For larger values of N_{real}, the jitter reduces, leading to converging curves in Fig. (3) (as mentioned earlier, it vanishes for $N_{real} \to \infty$).

In order to quantify the computational complexity of the proposed algorithm we define the relative execution time

$$R(N_{real}) = t_{GP-Draw}(N_{real})/t_{SP},$$

where $t_{GP-Draw}(N_{real})$ and t_{SP} denote the algorithm execution time for the same noisy speech material for *GP-Draw* and *SP*, respectively. Fig. 4 plots $R(N_{real})$ vs. N_{real}, and shows a relative computation time for *GP-Draw re. SP* in the range of 3.5-6 for $0 \le N_{real} \le 100$. Considering the complexity of a *entire* ASR system, note that the execution time for the ASR back-end (which is presumably the same for *GP-Draw* and

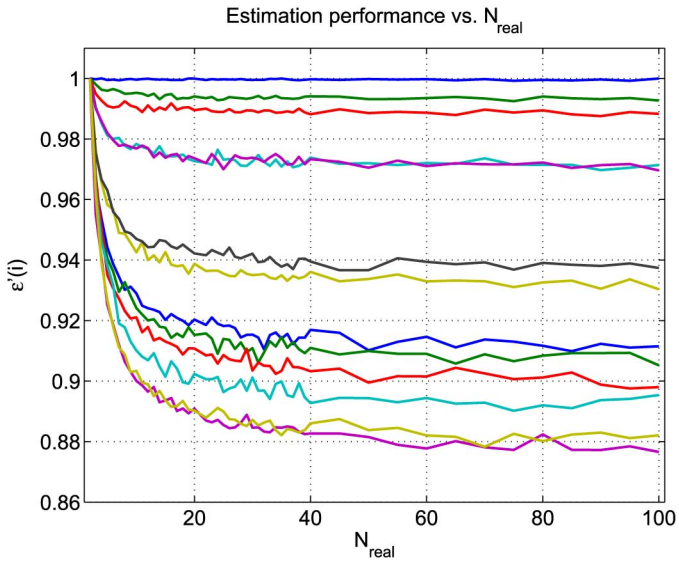

Fig. 3. Normalized MFCC mean-squared estimation error $\varepsilon'(i; N_{real})$ as a function of N_{real} for speech shaped noise and SNR $= 0$ dB, averaged across 75 noise realizations. The performance curves arrange themselves from top to bottom with increasing cepstral indices $i = 0, \cdots, I-1$, $I = 13$.

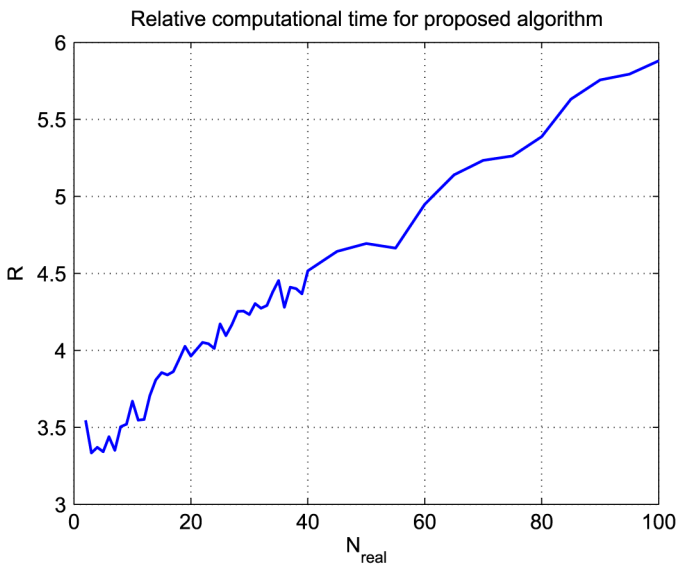

Fig. 4. Relative computation time of *GP-Draw re. SP* as a function of N_{real}.

SP) must be added to $t_{GP-Draw}(N_{real})$ and t_{SP}, respectively. For this reason, Fig. 4 represents the *worst case* relative computational time. For large-vocabulary ASR back-ends, the relative computational complexity could be significantly lower than shown in the figure. In our Matlab implementation of *SP*, execution time is approximately 1/20 times real-time.

ACKNOWLEDGMENT

The authors would like to thank three anonymous reviewers and the associate editor, whose constructive comments helped improve the presentation of this work.

REFERENCES

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

[2] Y. Chung and J. H. L. Hansen, "Compensation of snr and noise type mismatch using an environmental sniffing based speech recognition solution," *J. Audio, Speech, Music Process.*, vol. 2013.1, pp. 1–14, 2013.

[3] H. Xu, P. Dalsgaard, Z. H. Tan, and B. Lindberg, "Noise condition-dependent training based on noise classification and SNR estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2431–2443, Nov. 2007.

[4] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, pp. 733–736.

[5] D. Yu *et al.*, "Robust speech recognition using a cepstral minimum-mean-square-error motivated noise suppressor," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 1061–1070, Jul. 2008.

[6] A. Stark and K. Paliwal, "MMSE estimation of log-filterbank energies for robust speech recognition," *Speech Commun.*, vol. 53, pp. 403–416, 2011.

[7] K. M. Indrebo, R. J. Povinelli, and M. T. Johnson, "Minimum mean-squared error estimation of mel-frequency cepstral coefficients using a novel distortion model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1654–1661, Nov. 2008.

[8] A. Acero, L. Deng, T. Kristjansson, and J. Wang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP*, 2000.

[9] , Z.-H. Tan and B. Lindberg, Eds., *Automatic speech recognition on mobile devices and over communication networks*. London, U.K.: Springer-Verlag, Feb. 2008.

[10] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[11] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[12] R. M. Stern, "Applying physiologically-motivated models of auditory processing to automatic speech recognition," in *Proc. 3rd Int. Symp. Auditory Audiol. Res. (ISAAR)*, Aug. 2011.

[13] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 101–116, Sep. 2005.

[14] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.

[15] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP*, 2000, pp. 373–376.

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[17] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, USA, May 1996, vol. 2, pp. 629–633.

[18] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.

[19] P. C. Loizou, *Speech Enhancement: Theory and Practice.*. Boca Raton, FL, USA: CRC, 2007.

[20] Y. Ephraim and I. Cohen, , R. C. Dorf, Ed., "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook*, 3rd ed. Boca Raton, FL, USA: CRC, Taylor & Francis, 2006.

[21] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.

[22] I. Cohen, "Speech enhancement using super-Gaussian speech models and noncausal *a priori* SNR estimation," *Speech Commun.*, vol. 47, pp. 335–350, 2005.

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

[24] C.-P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.

[25] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1992.

[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991, Series in Communications.

[27] "ETSI," Speech processing, transmission and quality aspects (STQ), distributed speech recognition, advanced front-end feature extraction algorithm, compression algorithm, 2002, eS 202 050 v1.1.1.

[28] P. J. Wolfe and S. J. Godsill, "Simple alternatives to the Ephraim and Malah suppression rule for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 496–499.
[29] J. 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, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.
[30] DARPA, "TIMIT, Acoustic-Phonetic Continuous Speech Corpus," Oct. 1990, pp. 1–1.1, NIST Speech Disc.
[31] A. Varga and H. J. M. Steeneken, "Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–253, 1993.
[32] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[33] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4266–4269.
[34] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, Paris, France, 2000.
[35] S. J. Young *et al., HTK: Hidden Markov Model Toolkit V3.2.1, Reference Manual*. Cambridge, U.K.: Cambridge Univ. Speech Group, 2004.
[36] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1989, pp. 532–535.
[37] G. Hirsch and D. Pearce, "Applying the advanced ETSI frontend to the Aurora-2 task," Tech. Rep., 2006.
[38] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech, Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.
[39] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[40] T. V. Sreenivas, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 383–389, Sep. 1996.
[41] A. Kundu, S. Chatterjee, A. S. Murthy, and T. V. Sreenivas, "GMM based Bayesian approach to speech enhancement in signal/transform domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4893–4896.
[42] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel *a priori* SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4897–4900.
[43] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1993.
[44] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 39, no. 4, pp. 795–805, Apr. 1991.
[45] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.
[46] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement–A Survey of the State-of-the-Art*. San Rafael, CA, USA: Morgan and Claypool, 2013, ser. Synthesis Lectures on Speech and Audio Processing..

Jesper Jensen received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. student and Assistant Research Professor. From 2000 to 2007, he was a Post-Doctoral Researcher and Assistant Professor with Delft University of Technology, Delft, The Netherlands, and an External Associate Professor with Aalborg University. Currently, he is a Senior Researcher with Oticon A/S, Copenhagen, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is also a Professor with the Section for Multimedia Information and Signal Processing (MISP), Department of Electronic Systems, at Aalborg University. His main interests are in the area of acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.

Zheng-Hua Tan received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999. He is an Associate Professor in the Department of Electronic Systems at Aalborg University, Aalborg, Denmark, which he joined in May 2001. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA, an Associate Professor in the Department of Electronic Engineering at Shanghai Jiao Tong University, and a postdoctoral fellow in the Department of Computer Science at Korea Advanced Institute of Science and Technology, Daejeon, Korea. His research interests include speech and speaker recognition, noise-robust speech processing, multimedia signal and information processing, human–robot interaction, and machine learning. He has published extensively in these areas in refereed journals and conference proceedings. He is an Editorial Board Member/Associate Editor for Elsevier Computer Speech and Language, Elsevier Digital Signal Processing and Elsevier Computers and Electrical Engineering. He was a Lead Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He has served/serves as a program co-chair, area and session chair, tutorial speaker and committee member in many major international conferences.