Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling

Tobias May, Steven van de Par, and Armin Kohlrausch

Abstract—**Although the field of automatic speaker recognition (ASR) has been the subject of extensive research over the past decades, the lack of robustness against background noise has remained a major challenge. This paper describes a noise-robust speaker recognition system that combines missing data (MD) recognition with the adaptation of speaker models using a universal background model (UBM). For MD recognition, the identification of reliable and unreliable feature components is required. For this purpose, the signal-to-noise ratio (SNR) based mask estimation performance of various state-of-the art noise estimation techniques and noise reduction schemes is compared. Speaker recognition experiments show that the usage of a UBM in combination with missing data recognition yields substantial improvements in recognition performance, especially in the presence of highly non-stationary background noise at low SNRs.**

Index Terms—**Automatic speaker recognition (ASR), mask estimation, mel frequency cepstral coefficient (MFCC), missing data, noise robustness, universal background model (UBM).**

I. Introduction

WHEREAS single speaker recognition can be performed quite robustly in clean acoustic conditions, the recognition performance severely degrades in the presence of background noise [1]. The reduced performance is caused by the mismatch between the features which have been learned by the classifier under clean acoustic conditions and the features which are observed in adverse acoustic scenarios. Common approaches to reduce this mismatch are feature compensation methods such as cepstral mean subtraction [2] and relative spectral (RASTA) processing [3].

In contrast to compensating for the effect of environmental noise, a major step towards noise-robust speaker recognition is to modify the structure of the recognizer such that it only considers feature components which are believed to contain reliable information about the target signal. Considering a two-dimensional time–frequency (T–F) representation of a noisy speech signal, some T–F components will be dominated by the target signal, whereas other regions will be contaminated by background noise. In missing data (MD) recognition, the classification is performed only on that part of the observed spectro-temporal feature space that is believed to be reliable [4]. Two different approaches exist to deal with missing features. *Imputation* refers to the technique of replacing missing features with an estimate of the feature value, whereas *marginalization* basically ignores missing features. Marginalization has been shown to be superior to imputation in the context of speech recognition [4]. In order to perform marginalization, a mask is constructed which classifies the feature space into reliable and unreliable components. Like the spectral feature space, the mask is defined as a function of time and frequency. In the second step, the classification is solely based on the reliable feature components whereas the unreliable components are assumed to be masked by the background noise.

One major drawback of the MD framework is that it needs to be based on spectral features. This limitation is caused by the required correspondence between the mask and individual feature components. In contrast to spectral features, mel frequency cepstral coefficients (MFCCs) are orthogonalized by the discrete cosine transform (DCT) and therefore can be used for a more compact feature representation [5], where each cepstral feature is representing properties of the global spectral shape. As a result, cepstral features are more accurately modeled by recognizers which commonly assume independence of feature components (e.g., diagonal Gaussian mixture models). This may account for the better accuracy of cepstral-based recognition systems under clean acoustic conditions compared to MD-based recognition systems [4], [6]. It is also interesting to note that in the context of automatic speech recognition, the performance of an MD-based speech recognizer significantly decreases as the vocabulary size increases [7], which is generally less problematic for cepstral-based recognition systems. Possibly a similar effect may occur for speaker recognition, where the number of speakers which can be discriminated using the spectral feature representation may be limited due to the covariance between the feature components, which is not effectively modeled by a diagonal covariance matrix. Thus, whereas MD systems provide considerable advantages over cepstral-based techniques in terms of noise robustness, MD systems are limited by their inherent dependence on spectral features and therefore, a proper modeling of the speaker-dependent characteristics becomes especially important for MD-based recognition systems in order

Manuscript received May 07, 2010; revised November 05, 2010; accepted May 11, 2011. Date of publication May 31, 2011; date of current version November 09, 2011. This work was supported by Philips Research. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nestor Becerra Yoma.

T. May and S. van de Par are with the Institute of Physics, University of Oldenburg, 26111 Oldenburg, Germany (e-mail: tobias.may@uni-oldenburg.de).

A. Kohlrausch is with the Philips Research, NL-5656 AE Eindhoven, The Netherlands, and also with the Department of Human Technology Interaction, Eindhoven University of Technology, NL-5600 MB Eindhoven, The Netherlands.

Digital Object Identifier 10.1109/TASL.2011.2158309

to provide a substantial benefit over conventional MFCC-based recognizers.

Many state-of-the art speaker recognition systems approximate the speaker-dependent distribution of features by Gaussian mixture models (GMMs) [8]. GMM-based speaker models are predominantly used for MD-based speaker recognition [9]–[14]. In cepstral-based speaker recognition systems, the usage of an universal background model (UBM) in combination with GMMs is well established and was shown to outperform GMM-based speaker recognition [15]. A UBM represents the speaker-independent distribution of features and speaker models are obtained by adapting the well-trained UBM parameters to the speaker-dependent speech material. Despite its superior performance, the possible benefit of using a UBM in combination with MD-based speaker recognition has not been investigated.

In this study, we combine the UBM-based adaptation of speaker models with missing data recognition. It is expected that the representation of spectral features can be substantially improved by using a UBM model, especially for recognizing speakers for which there is a limited set of training material. Because the UBM is trained on the pooled speech material of many speakers, it is possible to significantly increase the number of Gaussian mixture components and therefore, develop a more precise model of the feature distribution without the risk of over-training. In order to show the potential benefit of combining missing data with the UBM-based adaptation of speaker models, a missing data mask is required that indicates whether a feature component is reliable or missing. Because the estimated mask is the most critical component in missing data systems that limits the overall recognition performance, an extensive comparison of methods for estimating the missing data mask based on the local signal-to-noise ratio (SNR) is performed. Therefore, various strategies for obtaining an estimation of the noise and the clean speech spectrum are systematically compared and evaluated in non-stationary noise conditions. In this way, the best performing method for the GMM-based missing data recognizer is found and will serve as a baseline for the newly proposed approach.

The remainder of the paper is organized as follows. Section II gives an overview about missing data classification, the adaptation of UBM-based speaker models and discusses various ways to derive an estimate of reliable feature components based on a local SNR criterion. Section III outlines the evaluation procedure and the baseline system. In Section IV, speaker recognition experiments are conducted to analyze the benefit of using a UBM in combination with MD recognition and to evaluate various mask estimation procedures. Section V summarizes the main findings and concludes the paper.

II. Automatic Speaker Recognition System

In this section, the missing data-based speaker recognition system and the adaptation of speaker models from a UBM are described. Furthermore, various methods will be presented for deriving the required missing data mask based on the spectral estimation of the noise and speech components.

A. *Missing Data Recognition Using Adapted Gaussian Mixture Models*

Gaussian mixture models are used to approximate the probability distribution of the D-dimensional feature vector $\vec{x}$ for the task of speaker recognition. Assuming K diagonal Gaussian mixture components, the probability density function (pdf) of a GMM is given by [8]

$$p(\vec{x}|\lambda) = \sum_{c=1}^{K} w_c \prod_{m=1}^{D} \mathcal{N}(x_m, \mu_{c,m}, \sigma_{c,m}^2) \tag{1}$$

where w_c is the component weight and $\mathcal{N}(x_m, \mu_{c,m}, \sigma_{c,m}^2)$ is a uni-variate Gaussian distribution with mean $\mu_{c,m}$ and variance $\sigma_{c,m}^2$

$$\mathcal{N}(x_m, \mu_{c,m}, \sigma_{c,m}^2) = \frac{1}{\sqrt{2\pi\sigma_{c,m}^2}} \exp\left(-\frac{(x_m - \mu_{c,m})^2}{2\sigma_{c,m}^2}\right). \tag{2}$$

The model for each specific speaker can be summarized by the following set of parameters

$$\lambda = \left(w_c, \vec{\mu}_c, \vec{\sigma}_c^2\right) \quad \forall\, c = 1, \ldots, K. \tag{3}$$

In missing data recognition, the feature vector $\vec{x}$ is split into two sub-vectors, according to reliable R and unreliable U components, and both are treated differently during the classification process. The evidence of the reliable feature components is directly used to estimate the likelihood of the speaker identity λ. Although the unreliable components are assumed to be dominated by additive noise, they do contain information about the maximum energy of the target speech component. The assumption that the unreliable feature components are bounded between zero and the observed spectral energy is exploited by bounded marginalization [4], where the average likelihood is computed across the range of all possible levels that the unreliable components might have had (also called *counter-evidence*)

$$\begin{aligned} p(\vec{x}|\lambda) = &\sum_{c=1}^{K} w_c \prod_{r \in R} \mathcal{N}(x_r, \mu_{c,r}, \sigma_{c,r}^2) \\ &\times \underbrace{\prod_{u \in U} \frac{1}{x_{\text{high},u} - x_{\text{low},u}} \int_{x_{\text{low},u}}^{x_{\text{high},u}} \mathcal{N}(x_u, \mu_{c,u}, \sigma_{c,u}^2) dx_u}_{\text{counter-evidence}}. \end{aligned} \tag{4}$$

The integral in (4) can be evaluated as the vector difference of error functions [4], and (4) can be rewritten as (5), as shown at the bottom of the next page. The bounds were set to $[x_{\text{low},u}, x_{\text{high},u}] = [0, x_u]$.

The speaker-dependent set of GMM parameters λ listed in (3) is commonly initialized by k-means clustering [16] and further refined using the expectation–maximization (EM) algorithm [17]. The objective of selecting the number of Gaussian components K is to find the minimum model complexity which is required to accurately model the characteristics of all speakers [8]. As discussed in the introduction, instead of estimating the GMM parameters for each speaker independently, a speaker-independent UBM is used, which is trained on the pooled speech material of many speakers using k-means clustering and the EM algorithm [15]. A speaker-dependent model is derived by adapting the well-trained UBM parameters to the speech material of the corresponding speaker using maximum *a posteriori* (MAP) estimation. During the adaptation process, only those Gaussian components of the UBM are adapted, which show sufficient probabilistic alignment with the speaker-dependent speech material. In this way, the parameters of Gaussian components which are potentially under-represented are not updated to the new data, making the model adaptation robust even to a small amount of training data [15]. The MAP adaptation was shown to outperform the estimation of GMM parameters using the maximum-likelihood (ML) approach [15].

B. Spectral Features

Spectral features are computed using the short-time Fourier transform (STFT) on a frame-by-frame basis. First, the input signal $x(n)$ is processed with a first-order pre-emphasis filter using a coefficient of 0.97 in order to enhance the spectral representation of high frequencies. This is a standard pre-processing technique for computing mel frequency cepstral coefficients, which is typically not applied if spectral features are extracted. However, it was found to be beneficial for missing data recognition in pilot experiments. Then, the signal is transformed into overlapping segments and the N-point STFT is computed as

$$X(i,k) = \sum_{n=0}^{N-1} w(n) x((i-1)L+n) \exp\left(\frac{-j2\pi kn}{N}\right) \quad (6)$$

where i indexes the frame number, k represents the frequency bin index corresponding to the frequency $f(k) = kf_s/N$, f_s specifies the sampling frequency, w is a Hamming window function, and L determines the frame shift in samples. To reduce the number of spectral components, the spectrum $X(i,k)$ is passed through an auditory filterbank that resembles the frequency resolution of the human auditory system, resulting in an auditory power spectrum

$$X_{\mathrm{FB}}^2(i,j) = \sum_{k=0}^{N-1} |h_{\mathrm{FB}}(k,j) \times X(i,k)|^2 \quad (7)$$

for $j = 1, 2, \ldots, M$, where $M = 32$ is the number of auditory filters and $h_{\mathrm{FB}}(k,j)$ is a matrix containing the frequency-dependent auditory filter weights. The center frequencies f_c of the auditory filterbank are equally distributed on the equivalent rectangular bandwidth (ERB) scale [18] using a spacing of 1 ERB between 80 Hz and 5000 Hz. The set of triangular auditory filter weights is computed as

$$h_{\mathrm{FB}}(k,j) = \begin{cases} 0, & \text{for } f(k) < f_c(j-1) \\ \frac{f(k)-f_c(j-1)}{f_c(j)-f_c(j-1)}, & \text{for } f_c(j-1) \le f(k) < f_c(j) \\ \frac{f(k)-f_c(j+1)}{f_c(j)-f_c(j+1)}, & \text{for } f_c(j) \le f(k) < f_c(j+1) \\ 0, & \text{for } f(k) \ge f_c(j+1). \end{cases} \quad (8)$$

In the following, the two-dimensional, time- and frequency-dependent auditory power spectrum will be referred to as T–F representation. Finally, the auditory power spectrum is loudness compressed by raising it to the power of 0.33 to obtain the spectral features which are used for recognition.

C. Mask Estimation

In order to perform missing data classification, a mask is required which classifies the T–F representation into reliable and unreliable components. The underlying concept of the mask is that T–F units are assumed to be reliable if they are dominated by the target source, whereas the unreliable T–F units are considered to be dominated by interfering noise. This formulation implies that the local SNR is known for individual T–F components. To establish an upper performance limit, it is common to employ an ideal binary mask (IBM), which assumes *a priori* knowledge about the local SNR [4]. It was shown that such an ideal binary mask yields excellent speech recognition performance [4] and can significantly increase speech intelligibility in multi-talker scenarios [19]. Therefore, the estimation of the IBM was suggested to be the main goal of computational auditory scene analysis (CASA) [20]. Various strategies have been proposed to estimate the IBM based on auditory grouping principles [21]–[23], binaural interaction [24]–[26] and assessing the local SNR [4], [9], [27]. It is, however, outside the scope of this paper to analyze and compare all existing approaches. Because the focus of the present study is to robustly identify speakers in the presence of noise, the mask $m(i,j)$ is determined by estimating the local SNR in individual T–F units. A local SNR criterion [4] of $\mathrm{LC} = 0$ dB is applied to decide whether a T–F unit is reliable

$$m(i,j) = \begin{cases} 1, & \text{if } 10\log_{10} \frac{\hat{S}_{\mathrm{FB}}^2(i,j)}{\hat{N}_{\mathrm{FB}}^2(i,j)} > \mathrm{LC} \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

The local SNR is obtained by comparing the estimated auditory power spectrum of speech $\hat{S}_{\mathrm{FB}}^2(i,j)$ to the estimated auditory power spectrum of noise $\hat{N}_{\mathrm{FB}}^2(i,j)$ in individual T–F units. The estimation of speech and noise components is carried out in the spectral domain before applying the auditory filterbank. This was shown to be superior to performing the spectral estimation

$$p(\vec{x}|\lambda) = \sum_{c=1}^{K} w_c \prod_{r \in R} \mathcal{N}(x_r, \mu_{c,r}, \sigma_{c,r}^2) \prod_{u \in U} \frac{1}{x_{\mathrm{high},u} - x_{\mathrm{low},u}} \frac{1}{2} \left[\mathrm{erf}\left(\frac{x_{\mathrm{high},u} - \mu_{c,u}}{\sqrt{2\sigma_{c,u}^2}} \right) - \mathrm{erf}\left(\frac{x_{\mathrm{low},u} - \mu_{c,u}}{\sqrt{2\sigma_{c,u}^2}} \right) \right] \quad (5)$$

within auditory bands [28]. After estimating the spectral magnitude of both speech $\hat{S}(i,k)$ and noise $\hat{N}(i,k)$, both spectra are transformed to the auditory domain in analogy to (7):

$$\hat{S}^2_{\mathrm{FB}}(i,j) = \sum_{k=0}^{N-1} \left| h_{\mathrm{FB}}(k,j) \times \hat{S}(i,k) \right|^2 \qquad (10)$$

$$\hat{N}^2_{\mathrm{FB}}(i,j) = \sum_{k=0}^{N-1} \left| h_{\mathrm{FB}}(k,j) \times \hat{N}(i,k) \right|^2. \qquad (11)$$

Because neither the noise nor the speech spectrum is generally known *a priori*, they need to be estimated based on the noisy signal spectrum $X(i,k)$. A plethora of methods exist to perform this task and the choice of both the noise estimation technique and the method to obtain an estimate of the clean speech spectrum can potentially influence the quality of the estimated IBM. However, only a few studies have investigated the effect of some basic noise estimation techniques on MD recognition performance [27], [29]. Since the estimated IBM is the most critical component in missing data recognition systems, we will describe the most important methods to derive an estimate of the noise and the speech spectrum and their influence on the estimated IBM will be assessed and systematically evaluated in terms of speaker recognition performance in Sections IV-B and IV-C. A brief explanation of the compared algorithms will be given in the following. For a detailed description, the reader is referred to the corresponding references.

1) Noise Spectrum $\hat{N}(i,k)$: The estimate of the noise spectrum is derived from the noisy signal spectrum $X(i,k)$. Various noise estimation techniques have been proposed to deal with stationary and fluctuating noise types. A comprehensive overview and implementational details can be found in [30]. In this study, the most relevant developments are compared in the context of speaker recognition. The most simple method *Initial50ms* is estimating the noise floor by averaging the spectrum of the initial frames [27], assuming that no speech is present. The weighted average method *Hirsch95* introduced by Hirsch and Ehrlicher is using a first order recursion and employs an adaptive threshold to stop the recursion when speech activity is detected [31]. An alternative approach *Lin03* is adjusting the first-order recursion based on the estimated *a posteriori* SNR [32]. More elaborated methods, such as *Doblinger95*, *Cohen02* and *Martin06*, recursively average the noise power by tracking minima in the noisy spectrum [33]–[36]. A modification *Rangachari06* that aims at reducing the adaptation time of the noise estimate especially for highly non-stationary conditions is using a smoothing factor based on speech presence probability [37], [38]. Finally, a modified version *Lin03Mod* of the SNR-dependent recursive averaging [32] was implemented by smoothing the noisy spectrum $X(i,k)$ with a first-order recursion prior to estimating the SNR-dependent smoothing parameter. The frame-based smoothing was performed with a filter coefficient of $\eta = 0.6$ and aimed at reducing the variance of the resulting noise estimate. In Table I, the parameter settings of all evaluated noise estimation techniques are listed. As far as possible, parameters were chosen according to the recommendations of the authors.

TABLE I
EVALUATED NOISE ESTIMATION TECHNIQUES

Method	Parameters[1]	Reference
Initial50ms	average spectrum over initial 50 ms	[27]
Hirsch95	first-order recursion, $\alpha = 0.9$, $\beta = 2.5$	[31]
Doblinger95	$\alpha = 0.8$, $\beta = 0.96$, $\gamma = 0.998$	[33]
Cohen02	same as in [34]	[34]
Lin03	$\alpha_{\mathrm{final}} = 0.9$, $Q = 2$	[32]
Lin03Mod	$\alpha_{\mathrm{final}} = 0.9$, $Q = 2$, $\eta = 0.6$	[32]
Martin06	same as in [36]	[36]
Rangachari06	same as in [38]	[38]

[1] The notation of parameters corresponds to the notation in the corresponding reference. Therefore, the same parameter may have a different meaning across references.

2) Speech Spectrum $\hat{S}(i,k)$: In addition to the estimated noise spectrum $\hat{N}(i,k)$, an estimate of the clean speech spectrum $\hat{S}(i,k)$ is required to construct the missing data mask according to (9). In the context of noise reduction, much effort has been directed at improving the perceived quality of the estimated speech signal by attenuating the amount of residual noise while keeping speech and background noise artifacts at a minimum. However, these perceptual constraints are not necessarily relevant for the estimation of the ideal binary mask and it is a question, if perceptual improvements can be quantified in terms of recognition performance. To answer this question, the most common approaches are briefly reported and the influence on the estimated ideal binary mask will be assessed in Section IV-C.

An estimate of the speech spectrum $\hat{S}(i,k)$ can be derived by subtracting the estimated noise spectrum $\hat{N}(i,k)$ from the corrupted signal spectrum $X(i,k)$. The most frequently used technique to accomplish this is to perform spectral subtraction [42] by applying an SNR-dependent gain function in the frequency domain. The residual noise that exhibits strong temporal fluctuations after processing is often referred to as *musical noise* [43]. To reduce this problem of musical noise in spectral subtraction-based noise reduction schemes, an over-estimation of the noise in combination with introducing a spectral floor was found to be beneficial [44]. In the context of detecting reliable feature components based on spectral subtraction, the optimal over-estimation factor was reported to be close to $\alpha = 3$ [10], [45]. An alternative approach is to estimate the optimal minimum mean square error (MMSE) short-time spectral amplitude (STSA), which was reported to significantly reduce the problem of musical noise by recursively smoothing the *a priori SNR* [40], [41]. More recently, model-based approaches [46]–[48] have been reported to further improve the performance of speech enhancement systems especially in the presence of non-stationary noise at the expense of increasing computational complexity, but bearing in mind that the mask estimation is part of the front-end for missing data classification, we limited the estimation of the clean speech spectrum to SNR-based gain functions, which can be efficiently applied in the spectral domain.

To study the effect of the above described approaches on the estimation of the clean speech spectrum, and consequently on the estimation of the ideal binary mask, the following four gain functions are evaluated: magnitude spectral subtraction, power spectral subtraction, MMSE STSA, and MMSE log-STSA. The

TABLE II
EVALUATED GAIN CURVES FOR ESTIMATING THE CLEAN SPEECH SPECTRUM

Gain function	Parameters[1]	Reference
SpecSubPowα_1	$\alpha = 1$, $\beta = 0.01$, $\gamma_1 = 1$, $\gamma_2 = 1$	[39]
SpecSubPowα_3	$\alpha = 3$, $\beta = 0.01$, $\gamma_1 = 1$, $\gamma_2 = 1$	[39]
SpecSubMagα_1	$\alpha = 1$, $\beta = 0.01$, $\gamma_1 = 2$, $\gamma_2 = 0.5$	[39]
SpecSubMagα_3	$\alpha = 3$, $\beta = 0.01$, $\gamma_1 = 2$, $\gamma_2 = 0.5$	[39]
MMSE STSA	$\alpha = 0.98$	[40]
MMSE log-STSA	$\alpha = 0.98$	[41]

[1] The notation of parameters corresponds to the notation in the corresponding reference. Therefore, the same parameter may have a different meaning across references.

gain functions were implemented using the speech processing toolbox *VOICEBOX* [49]. Parameters of all tested gain functions are listed in Table II. Note that due to the convention in the corresponding references, the parameter α has a different meaning within the spectral subtraction and the MMSE framework. Spectral subtraction-based gain functions are characterized by the over-estimation factor α, the spectral floor β and the two exponents γ_1 and γ_2 in the generalized spectral subtraction scheme, which define the suppression rule to be either magnitude or power spectral subtraction. The MMSE-based gain curves are computed using a smoothing constant α to recursively estimate the *a priori SNR*.

III. Evaluation Setup

A. Acoustic Mixtures

Speaker recognition performance was evaluated on a closed set of 34 speakers (18 males and 16 females) using the SSC database [50]. The database consists of 17.000 clean utterances, 500 utterances per speaker. The audio signals were down-sampled to a sampling frequency of $f_s = 16$ kHz. To ensure that the speech material used to train the UBMs is different from the material used for the speaker recognition experiments, speech files of all 34 speakers were randomly partitioned into two equal sized sets, each consisting of 250 sentences per speaker. The first half was used to train the speaker-independent but gender-dependent UBMs for all recognizers. From the second half of the SSC database (again comprising 250 sentences per speaker), a certain amount of speech files was randomly selected for the speaker recognition experiments presented in Section IV. This limitation of speech files is motivated by the fact that the amount of available speech material is often a constraint for practical applications. The amount of speech material involved in the training of speaker-dependent models (using either a conventional GMM recognizer or the UBM-based adaptation, see Section III-B) and the evaluation of speaker recognition accuracy is reported for each experiment, individually. To investigate the influence on speaker recognition performance, the amount of available speech files is systematically varied in Section IV-D. In general, for each speaker 70% of the available speech material was randomly selected to train the corresponding speaker model and the remaining 30% was used for evaluation.

Because the assessed recognition accuracy will depend to some extent on the random selection of the training and the testing material, results are reported as the mean speaker identification accuracy over a series of 20 simulations, each containing a new randomly selected set of speech files for training and testing. Speaker recognition experiments were performed on either the full set of 34 speakers or with a subset of 10 randomly selected speakers. For each of the 20 simulations, a new subset of 10 speakers was randomly selected from the set of 34 speakers. In the testing phase, utterances were digitally mixed at various SNRs with noise signals drawn from the NOISEX database [51]. Five different noise types were used for evaluation: factory noise, cockpit noise, speech babble, destroyer operation engine noise and car noise. The SNR was computed by comparing the A-weighted energy of the speech signal to the A-weighted energy of the noise signal. The A-weighting filter was applied to ensure that the SNR is adjusted predominantly within the frequency range that is relevant for speech. The design of the A-weighting filter was implemented according to [52]. To prevent that the energy of speech is underestimated due to silent parts, an energy-based voice activity detector (VAD) was used to only consider signal segments with relevant speech activity. A frame was considered to contain relevant speech activity, if its energy level was within 40 dB of the global maximum.

B. GMM and UBM Parameters

In this study, two different types of recognizers were used to model the speaker-dependent distribution of features. The first recognizer, denoted as *GMM*, was a conventional GMM[2] recognizer. Based on features extracted from the speaker-dependent speech material, speaker models were first initialized by 20 iterations of the k-means algorithm and further trained using the EM algorithm. The EM algorithm iteratively refines the model parameters λ by maximizing the likelihood of the resulting GMM. The stopping criterion of the EM algorithm was set to $1e^{-5}$ with a maximum of 300 iterations. The second recognizer, referred to as *GMM-UBM*, utilized two gender-dependent UBMs, reflecting the distribution of speaker-independent features for male and female speech, respectively. The training of both UBMs involved one half of the SSC database (see Section III-A) and was accomplished by 20 initial k-means iterations followed by the EM algorithm using a stopping criterion of $1e^{-5}$ with a maximum of 300 iterations. Speaker-dependent models were obtained by adapting the trained UBM parameters to the speaker-dependent speech material. Therefore, first the gender selection was performed by selecting the UBM which showed the higher probabilistic alignment with the speaker-dependent speech material. Second, as suggested by [15], only the mean vectors of the UBM were adapted using a relevance factor of 16. The adaptation was performed by ten iterations of the EM algorithm. Note that the two recognizers were used both within the MD framework and also to represent the MFCC-based feature vectors (see Section IV-E). Only speech material with relevant speech activity was included in the training stage of both systems by using the previously described VAD.

[2] The GMM modeling was performed using the NETLAB package [53].

C. Baseline System

A conventional robust speaker recognition system was trained using a 26-dimensional feature vector consisting of 13 static MFCC coefficients, including the 0th order coefficient, and first order temporal derivatives, so called *delta coefficients*. The static MFCC coefficients were computed using the RASTAMAT toolbox [54]. Parameters[3] were chosen to reproduce MFCC coefficients according to the hidden Markov models toolkit (HTK), which is a commonly used front-end for speaker recognition experiments. The delta coefficients were computed using the trend derived from linear regression over a window of five frames [55]. In a pilot experiment, it was observed that the additional use of second-order temporal derivatives, so called *acceleration* coefficients reduced speaker recognition performance at low SNRs and therefore, acceleration coefficients were not appended. Cepstral mean and variance normalization (CMVN) was applied for improved robustness, where the feature statistics are measured over the duration of one utterance [56], [57]. Compared to cepstral mean normalization (CMN) [2], CMVN substantially improved MFCC-based recognition performance.

IV. Experiments

A series of speaker recognition experiments was conducted. The first experiment investigated the benefit of combining MD recognition with the UBM-based adaptation of speaker models under idealized conditions, assuming that the required ideal binary mask is known *a priori*. In reality, the IBM is not known and needs to be estimated. Therefore, experiments two and three aimed at exploring various techniques to estimate the IBM in a variety of different background noise conditions. In addition, a detailed analysis of errors made by the mask estimation process is presented. Based on these findings, the best performing method for estimating the missing data mask was selected in the fourth experiment and the performance improvement with the UBM-based missing data recognizer as a function of available speech material is investigated. Finally, the fifth experiment compared speaker recognition performance of the UBM-based missing data recognizer with a GMM-based MD recognizer and with state-of-the art MFCC-based recognizers.

A. Experiment 1: Effect of UBM Using an IBM

The first experiment studied the effect of using a UBM within the MD framework. To isolate the effect of the UBM on speaker recognition performance, this initial experiment used the ideal binary mask for MD recognition. In order to further investigate the effect of erroneously classified T–F components on missing data recognition, the ideal binary mask was modified by randomly labeling unreliable T–F components as reliable. A spectral feature component corresponding to an unreliable T–F unit is dominated by background noise and thus is likely to cause a mismatch between the training and the testing situation. The number of randomly modified components was chosen to be 20% of the number of reliable T–F components. Speaker recognition accuracy was evaluated on a subset of 10 speakers. The amount of available speech material per speaker was limited to 25 sentences, using 18 for training and 7 for testing. The optimal model complexity of both recognizers, *MD GMM IBM* and *MD GMM-UBM IBM*, was individually selected based on pilot experiments. For both systems, a model complexity of 64 Gaussian components was chosen.

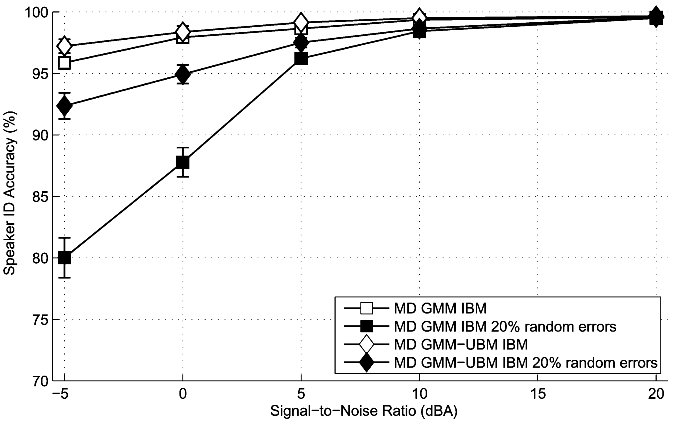

Fig. 1. Experiment 1: SNR-dependent speaker recognition performance for 10 speakers in the presence of factory noise using the ideal binary mask (IBM). The average recognition performance over a series of 20 simulations is presented for both recognizers, the GMM-based missing data system *MD GMM IBM* (squares) and the system including a universal background model *MD GMM-UBM IBM* (diamonds). The error bars represent the standard error of recognition performance across all 20 simulations.

The average SNR-dependent speaker recognition accuracy in the presence of factory noise is presented in Fig. 1. Note that the standard error of recognition performance across all 20 simulations was below 2% for all experimental conditions. Open symbols represent the IBM and black symbols indicate that the IBM has been modified by randomly labeling unreliable T–F units as being reliable. When using the IBM, the recognition performance of both systems *MD GMM IBM* and *MD GMM-UBM IBM* is almost identical. However, it can be seen that the MD recognizer in combination with a UBM is significantly more robust when unreliable T–F components are randomly labeled as reliable. Especially at low SNRs, the advantage of the UBM-based system in terms of speaker recognition accuracy is larger than 12%. When using the *MD GMM* system, all speaker-dependent model parameters are trained using the training data of one speaker only, and as a result, observations which have not been seen during the training stage can cause erroneous likelihood values which can potentially bias the class decision in very different ways across different speakers. Regarding the UBM-based recognition system, all speaker-dependent models share the same initial model parameters that have been learned based on the pooled speech material of the SSC database. Furthermore, only the mean values of the Gaussian components are adapted to the speaker-dependent training data that show sufficient probabilistic alignment. As a result, the UBM-based system is significantly less sensitive to observations which were not included in the training stage.

B. Experiment 2: Influence of Noise Estimation Algorithms

The second experiment investigated the influence of the noise floor prediction $\hat{N}(i,k)$ on the estimation of the IBM in

[3] melfcc(in, fs, 'lifterexp,' -22, 'nbands,' 20, 'dcttype,' 3, 'maxfreq,' 8000, 'fbtype,' 'htkmel,' 'sumpower,' 0, 'wintime,' 20e-3, 'hoptime,' 10e-3, 'numcep,' 13). For details, see http://labrosa.ee.columbia.edu/matlab/rastamat/mfccs.html

terms of speaker recognition performance. Independent of the noise estimation technique, the speech spectrum $\hat{S}(i,k)$ was obtained by applying the MMSE log-STSA gain function to the noisy input spectrum $X(i,k)$. Speaker models were trained using a GMM-based missing data recognizer *MD GMM* with 16 Gaussian components. Speaker recognition accuracy was evaluated on a subset of 10 speakers, involving a total of 25 sentences per speaker (18 sentences for training and 7 sentences for testing). The SNR-dependent recognition accuracy for all evaluated noise estimation techniques is presented in Table III. The upper five panels show recognition performance for different noise types, whereas the last table depicts the average performance over all noise conditions. The corresponding standard error of recognition performance across all 20 simulations was below 2% for all experimental conditions. Results are ranked for different noise types according to recognition performance, starting with the lowest performance for babble noise and ending with the highest performance in the presence of car noise. This ranking coincides with the stationarity of the background noise, ranging from very non-stationary (babble and factory noise) to more stationary conditions (car noise). At an SNR of 20 dBA, no substantial difference was observed between the evaluated noise estimation techniques. Considering lower SNRs, several methods performed well for one particular noise type but recognition performance was considerably lower compared to other methods if the speech material was corrupted by other types of background noise. *Rangachari06* performed well under the influence of both cockpit and destroyer noise, but recognition performance was significantly lower than other methods in the presence of factory, babble and car noise. *Doblinger95* and *Hirsch95* performed best in the condition of cockpit noise but performance was way below *Cohen02* and *Lin03Mod* in all other background noise scenarios. This sensitivity to the type of background noise suggests that the corresponding noise estimation methods are most suitable for certain types of background noise or that the corresponding parameters may have been optimized for a particular noise condition. On average, the recognition performance using *Martin06* was about 10% below *Lin03Mod* at low SNRs, which might have been caused by the rather conservative adaptation of the noise spectrum and the comparably long initialization phase. The overall lowest recognition performance was obtained by *Lin03*. This method was initially designed to estimate the noise spectrum in auditory bands. Clearly, the spectral variance of frequency bins is much larger than the variation within auditory bands and therefore, this method failed to successfully predict the noise power spectrum. In contrast to the above mentioned techniques, *Cohen02* and *Lin03Mod* were consistently among the best methods across all evaluated noise types. Compared to *Lin03*, the additional recursion in *Lin03Mod* significantly increased recognition performance.

TABLE III
EXPERIMENT 2: AVERAGE MISSING DATA SPEAKER RECOGNITION ACCURACY OVER A SERIES OF 20 SIMULATIONS FOR A SUBSET OF 10 SPEAKERS IN THE PRESENCE OF DIFFERENT TYPES OF BACKGROUND NOISE. THE IDEAL BINARY MASK (IBM) WAS ESTIMATED USING THE *MMSE log-STSA* NOISE REDUCTION SCHEME AND VARIOUS NOISE ESTIMATION TECHNIQUES LISTED IN TABLE I

babble noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	29.86	58.57	84.50	96.79	98.86
Hirsch95	20.43	45.00	76.50	94.93	98.86
Doblinger95	20.50	46.50	79.43	95.14	98.79
Cohen02	34.36	69.71	90.86	97.86	98.93
Lin03	16.29	36.50	66.29	90.64	98.86
Martin06	22.29	51.50	79.50	95.79	99.07
Rangachari06	23.07	50.86	82.57	96.07	98.93
Lin03Mod	33.07	68.29	91.50	97.71	99.00
factory noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	35.64	61.93	84.93	95.64	98.86
Hirsch95	23.00	45.86	75.43	93.57	99.07
Doblinger95	24.43	48.86	79.36	94.93	98.79
Cohen02	39.86	69.50	89.43	96.57	98.86
Lin03	16.43	32.64	60.57	86.57	98.86
Martin06	26.86	54.00	81.07	94.86	98.86
Rangachari06	30.29	61.29	87.00	96.57	98.64
Lin03Mod	39.29	70.93	88.86	96.07	98.79
destroyer noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	45.79	71.93	89.43	96.21	99.00
Hirsch95	24.57	49.00	79.00	94.29	99.00
Doblinger95	31.14	62.00	86.86	96.29	98.71
Cohen02	45.57	77.64	92.50	97.36	99.07
Lin03	13.93	29.07	56.86	84.57	98.71
Martin06	35.86	64.36	86.50	96.07	99.00
Rangachari06	55.57	82.50	94.64	97.50	98.93
Lin03Mod	50.79	77.14	91.36	96.79	99.00
cockpit noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	79.50	93.93	97.36	98.79	99.00
Hirsch95	77.57	93.29	97.36	98.64	99.07
Doblinger95	81.57	94.21	97.57	98.21	98.71
Cohen02	73.86	92.64	97.07	98.29	99.00
Lin03	26.00	57.43	85.14	94.79	98.79
Martin06	73.71	91.64	97.50	98.64	99.00
Rangachari06	81.71	94.93	97.14	98.07	98.86
Lin03Mod	83.50	93.64	96.71	97.93	98.93
car noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	72.07	85.00	91.43	95.57	98.14
Hirsch95	60.21	75.57	87.50	94.71	98.50
Doblinger95	60.64	77.07	87.64	95.50	97.64
Cohen02	80.50	92.93	96.29	97.57	98.50
Lin03	22.07	28.36	45.93	68.43	96.43
Martin06	67.29	80.29	87.86	93.29	98.29
Rangachari06	72.43	84.43	91.29	96.21	98.21
Lin03Mod	86.14	93.43	96.71	98.07	98.50
average	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
Initial50 ms	52.57	74.27	89.53	96.60	98.77
Hirsch95	41.16	61.74	83.16	95.23	98.90
Doblinger95	43.66	65.73	86.17	96.01	98.53
Cohen02	54.83	80.49	93.23	97.53	98.87
Lin03	18.94	36.80	62.96	85.00	98.33
Martin06	45.20	68.36	86.49	95.73	98.84
Rangachari06	52.61	74.80	90.53	96.89	98.71
Lin03Mod	58.56	80.69	93.03	97.31	98.84

In order to gain more understanding of the underlying factors that influence the recognition performance, receiver operating characteristics (ROC) curves [58] were computed by comparing the estimated IBM to the ideal binary mask which was computed using *a priori* knowledge of the speech and the noise spectra. The ROC graph visualizes the trade-off between correctly identified T–F components which are dominated by the target signal (true positive rate) and misclassified T–F elements which are dominated by background noise (false positive rate). The higher the true positive rate and the smaller the false positive rate, the higher the quality of the estimated IBM. Based on the experimental results reported in Table III, the corresponding SNR-dependent ROC curves are shown for all evaluated noise estimation techniques in Fig. 2. The ROC curves are averaged across all five background noise types. With decreasing SNR,

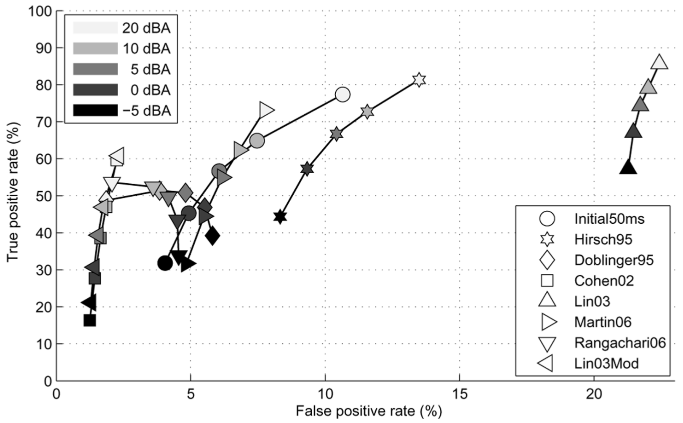

Fig. 2. Experiment 2: SNR-dependent ROC curves for all evaluated noise estimation algorithms. The ROC curves are averaged over all five noise conditions.

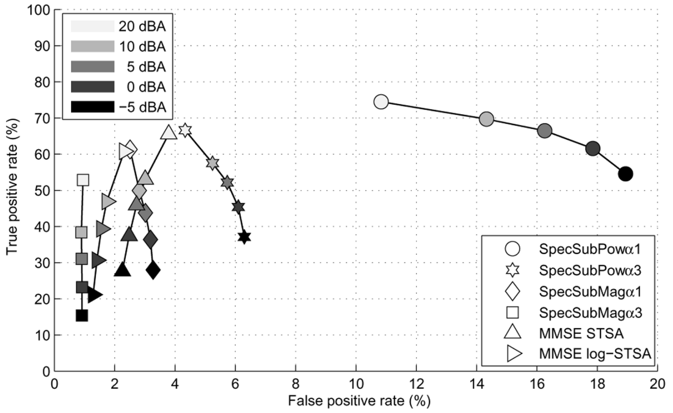

Fig. 3. Experiment 3: SNR-dependent ROC curves for all evaluated noise suppression rules. ROC curves are averaged over all five noise conditions.

the true positive rate decreased for all noise estimation techniques. Although *Lin03* achieved the highest true positive rates, the false positive rates are above 20% for all SNR conditions. The additional first-order recursion employed in *Lin03Mod* substantially reduced the false positive rate. It can be seen that the two best performing noise estimation methods *Lin03Mod* and *Cohen02* are most conservative in labeling reliable T–F components, keeping the false positive rate down to about 2%. The small advantage of *Lin03Mod* over *Cohen02* at low SNRs is manifested in a slightly higher true positive rate. Thus, it seems that false positive errors are most problematic for MD recognition using a GMM classifier.

C. Experiment 3: Influence of Speech Estimation Algorithms

In the third experiment, the impact of deriving an estimate of the clean speech spectrum $\hat{S}(i,k)$ on the estimated IBM was analyzed. The clean speech spectrum is estimated by subtracting the estimated noise floor from the noisy input spectrum using various SNR-based gain curves (see Table II). The objective of those gain curves is to enable maximum noise suppression while simultaneously minimizing the amount of speech distortion. The strength of the noise suppression determines the amount of noise energy which might leak into the estimated speech power spectrum, and consequently, will affect the amount of T–F elements in the estimated IBM which are erroneously identified as reliable components due to the over-estimation of the speech power spectrum.

The MD-based speaker recognition performance for the estimated IBM based on all tested methods to derive an estimate of the clean speech spectrum is presented in Table IV depending on the SNR and the type of background noise. The standard error of recognition performance across all 20 simulations was below 2% for all experimental conditions. The noise floor was estimated by *Lin03Mod*, which was the most consistent method among all evaluated noise estimation techniques in the second experiment. The remaining experimental conditions were identical to the second experiment (see Section IV-B). Similar to the second experiment, the SNR-dependent ROC curves corresponding to the experimental results in Table IV are presented in Fig. 3. Regarding spectral subtraction-based gain functions, magnitude-based spectral subtraction with an over-estimation factor of 3 $SpecSubMag\alpha_3$ enabled the strongest attenuation of the estimated noise and consequently, lowered the level of the estimated speech spectrum. Thus, $SpecSubMag\alpha_3$ conservatively selected reliable T–F components and as a result, achieved the highest recognition performance among all spectral subtraction-based algorithms. This observation is consistent across all evaluated background noise scenarios and can be confirmed by comparing the corresponding ROC curves, which show systematically decreasing false positive rates for noise reduction schemes with stronger noise attenuation. Although the MMSE-based gain functions *MMSE STSA* and *MMSE log-STSA* produced slightly higher false positive rates, *MMSE log-STSA* outperformed $SpecSubMag\alpha_3$ in terms of recognition accuracy, especially in conditions with low SNRs. Whereas moderate improvements are observed in the presence of highly nonstationary noise (babble and factory noise), more substantial benefits of *MMSE log-STSA* over the spectral subtraction-based methods are found for more stationary scenarios, namely destroyer and car noise. This advantage is presumably caused by the higher true positive rates of *MMSE log-STSA*, which effectively produced an estimated IBM which is less sparse than the mask obtained by $SpecSubMag\alpha_3$.

D. Experiment 4: Effect of UBM Using an Estimated IBM

The first experiment showed a significant benefit when using a UBM within the MD framework under ideal conditions (i.e., when the ideal binary mask is known *a priori*). Consequently, the fourth experiment aimed at verifying this benefit for scenarios in which the ideal binary mask is not known *a priori* and needs to be estimated. Based on findings of the second and third experiment, the IBM was estimated by combining the *Lin03Mod* noise estimation with the *MMSE log-STSA* gain function. In addition, the dependency of the reported performance gain on the amount of available speech material is investigated. The speech material that can be used for training and testing the speaker models consist of 250 sentences per speaker, which might not be available for practical applications. Furthermore, the amount of available speaker-dependent speech material and the complexity of the recognizer (number of Gaussian components used by the recognizer) are interdependent and both parameters are expected to influence the comparison between

TABLE IV
EXPERIMENT 3: AVERAGE MISSING DATA SPEAKER RECOGNITION ACCURACY OVER A SERIES OF 20 SIMULATIONS FOR A SUBSET OF 10 SPEAKERS IN THE PRESENCE OF DIFFERENT TYPES OF BACKGROUND NOISE. THE IDEAL BINARY MASK (IBM) WAS ESTIMATED USING THE *Lin03Mod* NOISE ESTIMATION TECHNIQUE AND VARIOUS NOISE REDUCTION SCHEMES LISTED IN TABLE II

babble noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	14.64	31.36	61.14	90.71	98.57
SpecSubPowα_3	19.21	42.50	75.86	95.29	99.14
SpecSubMagα_1	21.93	49.57	83.79	96.79	99.00
SpecSubMagα_3	32.07	67.29	90.71	97.57	99.00
MMSE STSA	28.57	60.36	88.43	97.21	99.00
MMSE log-STSA	33.07	68.29	91.50	97.71	99.00
factory noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	14.86	28.86	57.71	88.21	98.86
SpecSubPowα_3	18.93	42.86	77.93	94.79	99.21
SpecSubMagα_1	24.93	54.43	84.14	95.93	98.93
SpecSubMagα_3	35.71	69.57	89.21	96.29	98.71
MMSE STSA	33.43	64.71	86.36	96.07	98.93
MMSE log-STSA	39.29	70.93	88.86	96.07	98.79
destroyer noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	15.29	25.50	51.86	85.07	98.93
SpecSubPowα_3	18.50	37.93	71.79	93.29	99.14
SpecSubMagα_1	21.50	46.86	79.93	94.50	98.93
SpecSubMagα_3	33.79	65.79	86.71	95.64	98.50
MMSE STSA	45.36	74.29	90.29	96.86	99.14
MMSE log-STSA	50.79	77.14	91.36	96.79	99.00
cockpit noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	19.21	41.93	76.57	93.50	98.93
SpecSubPowα_3	47.50	82.21	95.64	98.14	99.00
SpecSubMagα_1	67.57	91.79	96.57	97.93	99.00
SpecSubMagα_3	76.64	91.00	95.43	97.29	98.71
MMSE STSA	86.71	94.71	97.14	98.29	99.07
MMSE log-STSA	83.50	93.64	96.71	97.93	98.93
car noise	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	16.57	18.50	30.50	52.14	93.71
SpecSubPowα_3	25.50	39.93	55.71	80.50	98.43
SpecSubMagα_1	32.36	54.86	72.50	91.86	98.71
SpecSubMagα_3	71.07	88.14	94.86	97.93	98.36
MMSE STSA	82.93	91.71	96.43	98.36	98.64
MMSE log-STSA	86.14	93.43	96.71	98.07	98.50
average	-5 dBA	0 dBA	5 dBA	10 dBA	20 dBA
SpecSubPowα_1	16.11	29.23	55.56	81.93	97.80
SpecSubPowα_3	25.93	49.09	75.39	92.40	98.99
SpecSubMagα_1	33.66	59.50	83.39	95.40	98.91
SpecSubMagα_3	49.86	76.36	91.39	96.94	98.66
MMSE STSA	55.40	77.16	91.73	97.36	98.96
MMSE log-STSA	58.56	80.69	93.03	97.31	98.84

both speaker recognition systems *MD GMM* and *MD GMM-UBM*. Therefore, the influence of the following parameters is analyzed:

- number of speaker-dependent sentences: 13, 25, 50, 125, and 250 of the SSC database (70% for training and 30% for testing)
- number of GMM components: 16, 32, 64, 128, and 256,
- type of recognizer: *MD GMM* and *MD GMM-UBM*.

The relative difference in speaker recognition accuracy between *MD GMM-UBM* and *MD GMM* is presented in the left panels of Fig. 4 as a function of the SNR and the number of Gaussian components. In addition, the right panels show the corresponding absolute speaker recognition accuracy using the *MD GMM-UBM* recognizer. Speaker recognition was performed on a subset of 10 speakers in the presence of factory noise. The amount of available speech material is systematically increased, ranging from 13 sentences in the top panels to 250 sentences in the bottom panels. It can be observed that the benefit of the *MD UBM-GMM* is greatest when a limited amount of speech material is available for training. In general, rather moderate improvements are observed for conditions with a high SNR down to 10 dBA, because the corresponding recognition performance is close to 100%, but considerable improvements are achieved for low SNRs. If only 13 sentences of speaker-dependent material are involved in the speaker recognition experiments as depicted in panel (a), the improvement can be as high as 23% for conditions with low SNRs. With increasing availability of speech material, the difficulty of properly training speaker-dependent GMMs decreases and as a consequence, the relative benefit of the UBM-based missing data system decreases. When all 250 sentences per speaker are used [see panel (e)], the benefit of applying a UBM is only about 10% at low SNRs. Overall, the usage of a UBM in conjunction with MD recognition shows a consistent and substantial benefit, especially in challenging noise conditions with low SNR.

The optimal number of Gaussian components moderately depends on the amount of available training material. If the amount of speech material is limited (to up to 25 sentences), the *MD GMM-UBM* recognizer with 64 Gaussian components performed best. This is consistent with the observation that the UBM-based recognizer with 64 components using the ideal binary mask performed best in the first experiment (see Section IV-A). The use of Gaussian mixture models with a higher complexity led to a decrease in recognition performance [see panels (f) and (g)], which may indicate that the resulting speaker models were overtrained given the limited amount of speech material. Having access to at least 50 sentences of the SSC database, the recognizer with 128 Gaussian mixtures slightly outperformed systems with lower complexity at low SNRs. Speaker recognition performance saturated at a model complexity of 128 Gaussian components and a further increase did not lead to significant improvements.

E. Experiment 5: MD Recognition Versus MFCCs

The last experiment compared the proposed UBM-based missing data recognizer with a conventional MFCC-based recognition system (see Section III-C). The distribution of both spectral and MFCC features was learned by the two classifiers *GMM* and *GMM-UBM* (see Section III-B). For MD recognition, the required mask was estimated using the *Lin03Mod* noise estimation technique and the *MMSE log-STSA* gain function. Furthermore, MD recognition was performed using the ideal binary mask in order to show the theoretical upper performance limit. In addition to the conventional MFCC-based recognizer, another baseline system is used that employed a noise reduction (NR) stage prior to computing the MFCC feature vector. Among all evaluated methods, the *Hirsch95* noise estimation technique in combination with the *MMSE log-STSA* gain function performed best during initial tests. This pre-processing is indicated by *NR*. Furthermore, analogously to the ideal binary mask, an ideal noise reduction front-end *NR IDEAL* was applied that used *a priori* knowledge about the noise floor. Similar to the first experiment, the optimal model complexity of each recognizer was selected based on pilot experiments. Whereas the two MFCC-based recognizers *MFCC GMM* and *MFCC*

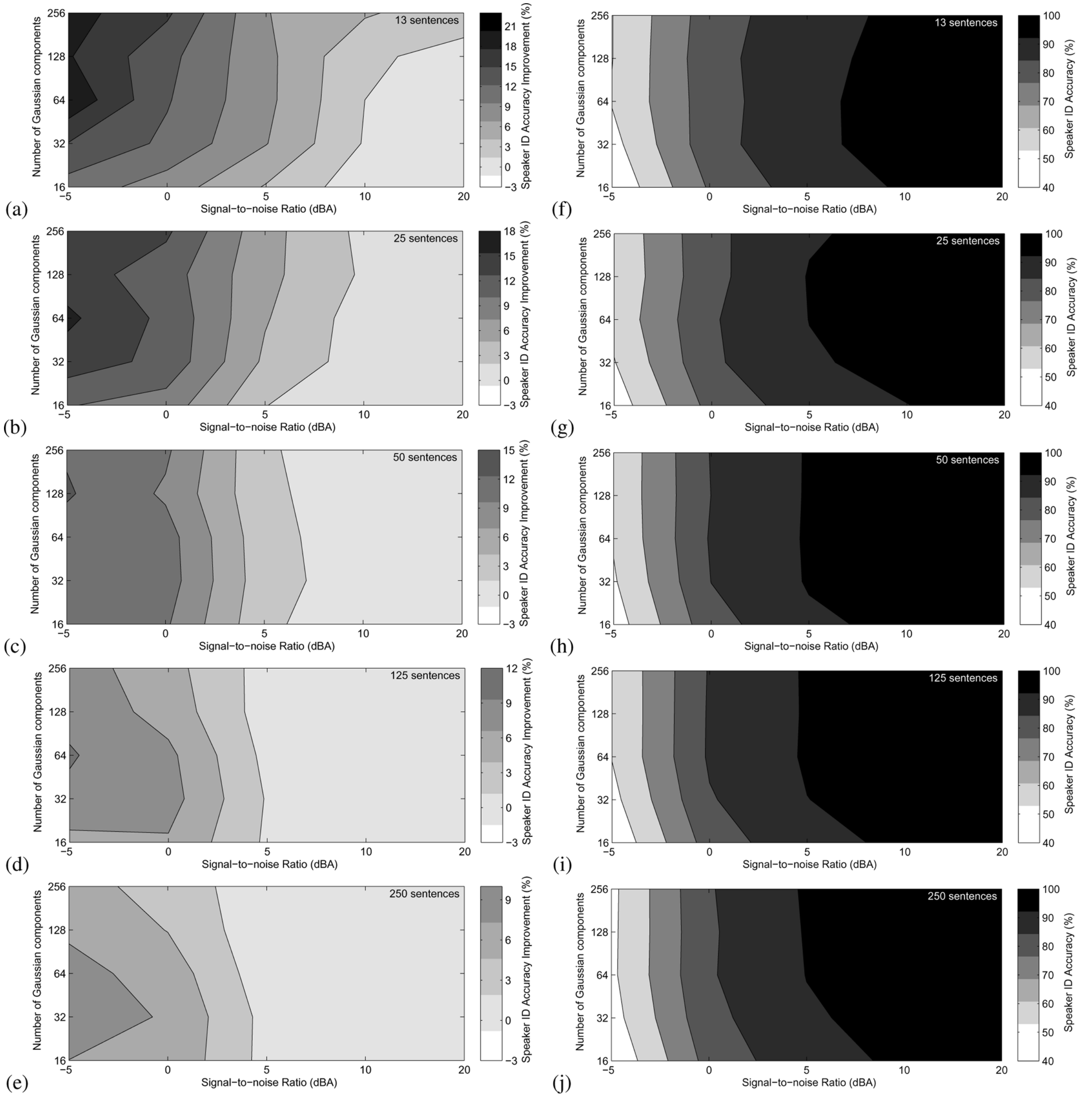

Fig. 4. Experiment 4: Speaker identification improvement of the *MD GMM-UBM* method compared to the *MD GMM* system on a closed set of 10 speakers in the presence of factory noise. The left panels show the speaker identification improvement as a function of the SNR and the number of Gaussian components for experiments involving (a) 13, (b) 25, (c) 50, (d) 125, and (e) 250 sentences of speaker-dependent speech material. The right panels (f)-(j) show the corresponding speaker identification accuracy of the *MD GMM-UBM* recognizer. Both speaker identification improvement and speaker identification accuracy are reported as the mean over a series of 20 simulations. The standard error of both measures was below 3% for all experimental conditions.

GMM NR performed best with 32 Gaussian components, all other recognizers utilized 128 Gaussian components.

The average speaker recognition performance is shown in Fig. 5 depending on the SNR. The left panels (a) and (b) represent speaker recognition performance on a subset of 10 speakers, whereas the right panels (c) and (d) show results for the full set of 34 speakers. For each speaker, a total of 25 sentences were used (18 sentences for training and 7 sentences for testing). The standard error across all 20 simulations was below 2% for all experimental conditions. Black symbols signify recognizers that are based on UBM adaptation. According to the results obtained in the second and third experiment, the five background noise types are grouped into two categories, namely highly non-stationary (babble and factory noise) and more stationary scenarios (destroyer, car, and cockpit noise). Speaker recognition performance is separately shown as the average of highly non-stationary (top panels) and more stationary noise conditions (bottom panels).

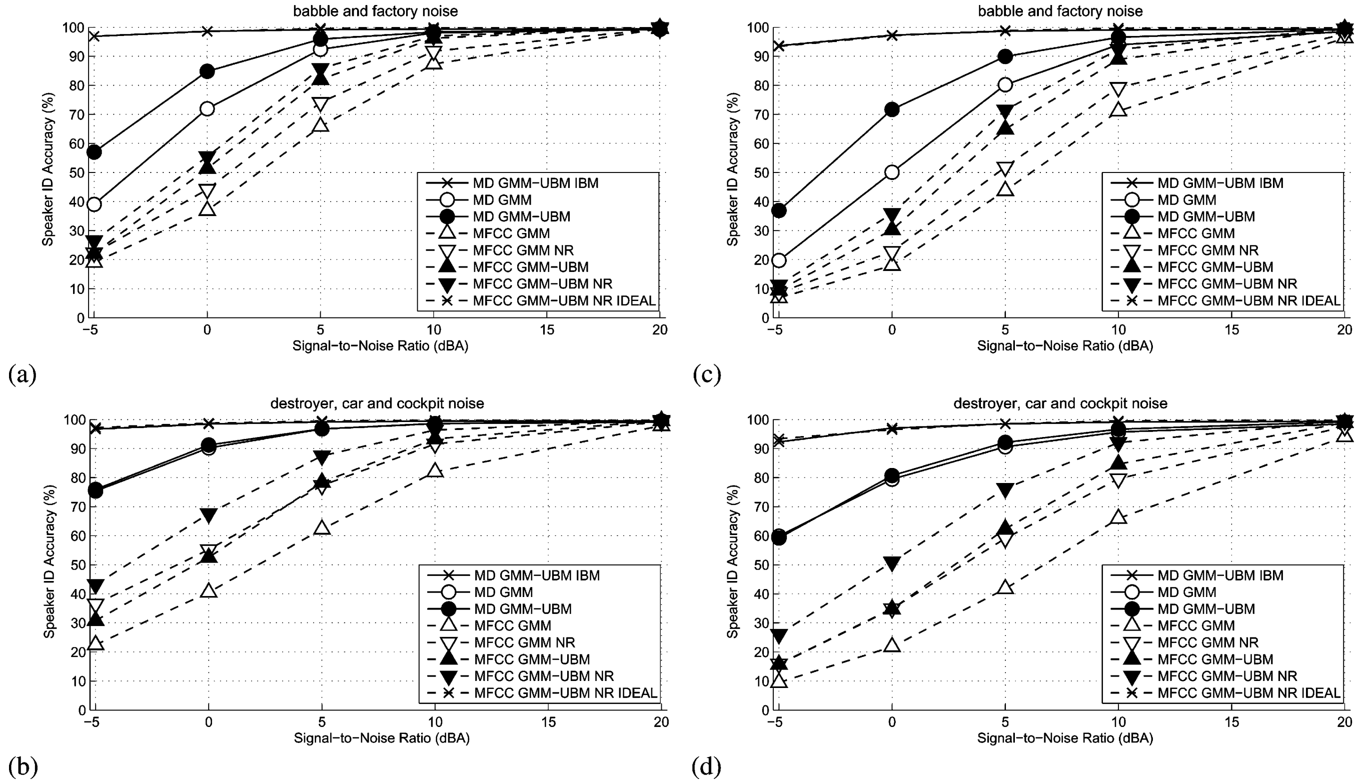

Fig. 5. Experiment 5: SNR-dependent speaker recognition performance on sets consisting of 10 speakers [panels (a) and (b)] and 34 speakers [panels (c) and (d)], respectively. Results are shown for two groups of background noise scenarios; highly non-stationary [panels (a) and (c)] and more stationary noise conditions [panels (b) and (d)]. Recognition performance is presented as the average recognition performance over a series of 20 simulations. The standard error of the reported recognition performance across all 20 simulations was below 2% for all experimental conditions. Results are presented for two types of recognizers, the MFCC-based recognizers (dashed lines) and the MD recognizers (solid lines). Black symbols indicate that the corresponding recognizers are based on UBM adapation. The two recognizers marked by crosses utilize *a priori* information about the noise floor.

Compared to the GMM-based cepstral recognizer *MFCC GMM* (△), the additional use of a UBM (▲) significantly improved recognition performance in all conditions, but already at moderate SNRs, the speaker recognition performance of both MFCC-based recognizers *MFCC GMM* and *MFCC GMM-UBM* rapidly decreased. Considering more stationary noise conditions, a consistent and significant performance gain was obtained when noise reduction was performed prior to computing the MFCC features (▼). However, this benefit is noticeably reduced in highly non-stationary conditions, probably because of the inability of the noise estimation technique to quickly adapt to sudden changes in noise level. In turn, MD-based speaker recognition was superior to all MFCC-based recognition systems. In the presence of highly non-stationary noise (babble and factory noise), the usage of an UBM in combination with MD recognition (•) substantially improved recognition performance, especially at low SNRs. The relative performance improvement of the *MD GMM-UBM* recognizer over the *MD GMM* system (○) was in the range of 20% at very low SNRs. This improvement was found for both sets consisting of 10 and 34 speakers. Whereas the benefit of the UBM for the MFCC-based recognizer decreased at lower SNRs, the improvement for the MD recognizer in highly non-stationary noise conditions increased with decreasing SNR. Regarding the more stationary noise types, the benefit of the UBM was not significant. One possible explanation might be that for more stationary noise types, the false positive rate of the estimated binary mask is substantially below the error rates obtained in fluctuating noise scenarios. As a result, a larger amount of T–F units is erroneously labeled as speech for fluctuating noise, increasing the mismatch between training and testing. Compared to the GMM-based MD recognizer *MD GMM*, the UBM-based MD recognizer *MD GMM-UBM* is not as sensitive to this mismatch, providing a benefit especially for non-stationary conditions.

Considering the *MD GMM-UBM* system in the presence of both highly non-stationary and more stationary noise scenarios, the speaker recognition performance was close to that of the recognizer using the ideal binary mask *MD GMM-UBM IBM* for SNRs as low as 5 dBA, which suggests that the *Lin03Mod* noise estimation combined with the *MMSE log-STSA* gain function produces high quality missing data masks. Comparing the overall recognition performance between the two sets consisting of 10 and 34 speakers, it can be noticed that the speaker recognition accuracy is lower for the larger set of speakers, especially at lower SNRs. A similar dependency was observed in the context of speech recognition, where the recognition accuracy decreased when the vocabulary size was increased [7].

The fifth experiment also showed the fundamental limitation of noise reduction schemes to improve speaker recognition performance of MFCC-based recognizers in the presence of noise. As long as *a priori* information about the noise floor is available, the MFCC recognizer including the ideal noise reduction front-end *MFCC GMM-UBM NR IDEAL* shows excellent recognition performance and is comparable to the MD system using the ideal binary mask *MD GMM-UBM IBM*. However, for MFCC-based recognition, there is a tremendous difference between the system that uses the ideal noise floor and the one that employs the best operating front-end for noise floor estimation *MFCC GMM-UBM NR*. This performance difference between the idealized and the realistic scenario is significantly smaller for the MD recognizer, although both systems employ a similar front-end for noise floor estimation. We believe that this performance gap can be explained by the conceptual difference between both recognizers and the way knowledge about the noise floor is incorporated in both systems. Whereas the MD system only requires a binary decision about the reliability of individual T–F units, the noise reduction front-end needs an accurate estimation of the instantaneous SNR in order to design the noise reduction filter, which is obviously much more difficult. Furthermore, errors in the estimated noise floor will have different effects on speaker recognition performance for both systems. Whereas, e.g., an over-estimation of the noise will cause the estimated binary mask to be more sparse, leaving fewer T–F elements for classification, the noise reduction front-end will attenuate speech which can potentially distort the resulting MFCC feature vector. Our experimental results suggest that the errors induced by the noise estimation are more problematic for MFCC-based recognizers.

V. Discussion and Conclusion

A robust speaker recognition system was presented, which combines missing data recognition with the adaptation of speaker models using universal background models. Compared to a GMM-based recognizer, the additional use of a UBM was shown to be especially beneficial in representing the spectral features in highly non-stationary noise conditions. The improvement was found to depend on the amount of available training material and was greatest for a small amount of speech material, which is often a constraint for practical applications.

The first experiment revealed that a MD recognizer in combination with a UBM was significantly more robust against false positive rates in the ideal binary mask and showed superior speaker recognition accuracy compared to a conventional GMM-based missing data recognizer. This benefit could be confirmed in the fourth experiment for conditions, where the IBM was estimated. One possible explanation is that the speaker models of a conventional GMM-based recognizer are initialized and trained independently. As a result, observations which were not well represented in the training stage can bias the likelihood computation across speaker models in very different ways. This problem is more likely to occur if only a small amount of training material is available or if T–F elements in the estimated IBM are erroneously classified as reliable elements. When using a UBM, all speaker models are equally initialized using the UBM parameters and only those Gaussian components are adapted to the speaker-dependent speech material that show sufficient probabilistic alignment. Thus, the risk of a biased likelihood computation is reduced.

In the context of estimating the ideal binary mask using a local SNR criterion, several noise estimation techniques and gain functions were evaluated. Substantial speaker recognition accuracy differences were found across methods. Among all tested noise estimation techniques, the minima-controlled recursive averaging [34] and a modified version of the SNR-dependent recursive averaging [32] consistently achieved the highest recognition performance across a variety of background noise scenarios. In combination with the aforementioned noise estimation techniques, the *MMSE log-STSA* gain function clearly outperformed spectral subtraction-based methods in terms of recognition accuracy in all noise conditions. The quality of the estimated IBM was further analyzed by ROC statistics. This analysis revealed that best results are obtained by a conservative labeling of reliable T–F components, which is reflected in a low false positive rate.

Due to the sparsity of the estimated IBM especially in conditions with low SNR, the speaker identification is only based on a fraction of the overall feature space. Further improvements can be expected if the decision about reliable and unreliable feature components is softened by using a fuzzy mask [59]. The current work focused on the recognition of speakers in noisy environments. Future work will also investigate the performance of the proposed system in the presence of reverberation and in scenarios with multiple target speakers.

Acknowledgment

We would like to thank Dr. A. Härmä and Dr. S. Srinivasan for helpful discussions and their feedback on an earlier version of this manuscript.

References

[1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, pp. 1–15, 1997.

[2] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.

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

[4] 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.

[5] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[6] K. J. Palomäki, G. J. Brown, and J. P. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Commun.*, vol. 43, no. 1–2, pp. 123–142, 2004.

[7] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1486–1501, 2006.

[8] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[9] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environment with combined spectral subtraction and missing data theory," in *Proc. ICASSP*, 1998, vol. 1, pp. 121–124.

[10] A. Drygajlo and M. El-Maliki, "Use of the generalized spectral subtraction and missing feature compensation for robust speaker verification," in *Proc. RLA2C*, 1998, pp. 80–83.

[11] Y. Shao and D. L. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proc. ICASSP*, 2003, vol. 2, pp. 205–208.

[12] Y. Shao and D. Wang, "Robust speaker recognition using binary time-frequency masks," in *Proc. ICASSP*, 2006, pp. 645–648.

[13] M. Kühne, D. Pullella, R. Togneri, and S. Nordholm, "Towards the use of full covariance models for missing data speaker recognition," in *Proc. ICASSP*, 2008, pp. 4537–4540.

[14] D. Pullella, M. Kühne, and R. Togneri, “Robust speaker identification using combined feature selection and missing data recognition,” in *Proc. ICASSP*, 2008, pp. 4833–4836.

[15] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[16] S. Lloyd, “Least squares quantization in PCM,” *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[17] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood estimation from incomplete data via the EM algorithm,” *J. R. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[18] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hear. Res.*, vol. 47, no. 1–2, pp. 103–138, Aug. 1990.

[19] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, “Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation,” *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[20] D. L. Wang, “On ideal binary masks as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, ch. 12, pp. 181–197.

[21] G. Hu and D. L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[22] G. Hu and D. L. Wang, “Auditory segmentation based on onset and offset analysis,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 396–405, Feb. 2007.

[23] N. Ma, P. Green, J. Barker, and A. Coy, “Exploiting correlogram structure for robust speech recognition with multiple speech sources,” *Speech Commun.*, vol. 49, pp. 874–891, 2007.

[24] N. Roman, D. L. Wang, and G. J. Brown, “Speech segregation based on sound localization,” *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, Oct. 2003.

[25] K. J. Palomäki, G. J. Brown, and D. L. Wang, “A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation,” *Speech Commun.*, vol. 43, no. 4, pp. 361–378, 2004.

[26] S. Harding, J. Barker, and G. Brown, “Mask estimation for missing data speech recognition based on statistics of binaural interaction,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 58–67, Jan. 2006.

[27] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, “Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study,” in *Proc. Eurospeech*, 1999, pp. 2407–2410.

[28] J. A. N. Flores and S. Young, “Continuous speech recognition in noise using spectral subtraction and HMM adaptation,” in *Proc. ICASSP*, 1994, vol. 1, pp. 409–412.

[29] C. Ris and S. Dupont, “Assessing local noise level estimation methods: Application to noise robust ASR,” *Speech Commun.*, vol. 34, pp. 141–158, 2001.

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

[31] H. Hirsch and C. Ehrlicher, “Noise estimation techniques for robust speech recognition,” in *Proc. ICASSP*, 1995, vol. 1, pp. 153–156.

[32] L. Lin, W. Holmes, and E. Ambikairajah, “Subband noise estimation for speech enhancement using a perceptual wiener filter,” in *Proc. ICASSP*, 2003, vol. 1, pp. 80–83.

[33] G. Doblinger, “Computationally efficient speech enhancement by spectral minima tracking in subbands,” in *Proc. Eurospeech*, 1995, pp. 1513–1516.

[34] I. Cohen and B. Berdugo, “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[35] 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.

[36] R. Martin, “Bias compensation methods for minimum statistics noise power spectral density estimation,” *Signal Process.*, vol. 86, no. 5, pp. 1215–1229, 2006.

[37] S. Rangachari, P. Loizou, and Y. Hu, “A noise estimation algorithm with rapid adaptation for highly nonstationary environments,” in *Proc. ICASSP*, 2004, pp. 305–308.

[38] S. Rangachari and P. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *Speech Commun.*, vol. 48, pp. 220–231, 2006.

[39] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” *IEEE Trans. Speech Audio Process*, vol. 7, no. 2, pp. 126–137, 1999.

[40] 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.

[41] 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.

[42] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[43] O. Cappe, “Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[44] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Proc. ICASSP*, 1979, vol. 4, pp. 208–211.

[45] P. Renevey and A. Drygajlo, “Statistical estimation of unreliable features for robust speech recognition,” in *Proc. ICASSP*, 2000, vol. 3, pp. 1731–1734.

[46] 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.

[47] 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.

[48] D. Y. Zhao, W. B. Kleijn, A. Ypma, and B. de Vries, “Online noise estimation using stochastic-gain HMM for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 835–846, May 2008.

[49] M. Brookes, “VOICEBOX: Speech Processing Toolbox for MATLAB,” 2009 [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[50] M. Cooke and T.-W. Lee, “Speech Separation and Recognition Competition,” 2006 [Online]. Available: http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm

[51] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, “The NOISEX-92 study on the effect of additive noise on automatic speaker recognition,” Speech Research Unit, Defence Research Agency, Malvern, U.K., Tech. Rep., 1992.

[52] *American National Standard Specification for Sound Level Meters*, ANSI/ASA S1.4-1983 (R2001), 1983, American National Standards Institute.

[53] I. T. Nabney and C. M. Bishop, NETLAB Package, 2001–2004, 2004 [Online]. Available: http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/

[54] D. P. W. Ellis, PLP and RASTA (and MFCC, and Inversion) in Matlab, 2009 [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[55] F. K. Soong and A. E. Rosenberg, “On the use of instantaneous and transitional spectral information in speaker recognition,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 6, pp. 871–879, Jun. 1988.

[56] J. P. Openshaw and J. S. Mason, “On the limitations of cepstral features in noise,” in *Proc. ICASSP*, 1994, pp. 49–52.

[57] S. Tibrewala and H. Hermansky, “Multi-band and adaptation approaches to robust speech recognition,” in *Proc. Eurospeech*, 1997.

[58] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recog. Lett.*, vol. 27, pp. 861–874, 2006.

[59] 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.

Tobias May received the Dipl.-Ing. (FH) degree in hearing technology and audiology from the Oldenburg University of Applied Science, Oldenburg, Germany, in 2005 and the M.Sc. degree in hearing technology and audiology from the University of Oldenburg, Oldenburg, Germany in 2007. He is currently pursuing the Ph.D. degree at the University of Oldenburg.

Since 2007, he has been with the Eindhoven University of Technology, Eindhoven, The Netherlands. Since 2010, he has been affiliated with the University of Oldenburg. His research interests include computational auditory scene analysis, binaural signal processing, and automatic speaker recognition.

Steven van de Par studied physics at the Eindhoven University of Technology, Eindhoven, The Netherlands, and received the Ph.D. degree from the Eindhoven University of Technology in 1998 on a topic related to binaural hearing.

As a Postdoctoral Researcher at the Eindhoven University of Technology, he studied auditory-visual interaction and was a Guest Researcher at the University of Connecticut Health Center. In early 2000, he joined Philips Research, Eindhoven, to do applied research in digital signal processing and acoustics. His main fields of expertise are auditory and multisensory perception, low-bit-rate audio coding and music information retrieval. He has published various papers on binaural auditory perception, auditory-visual synchrony perception, audio coding and music information retrieval (MIR)-related topics. Since April 2010, he has been a Professor in acoustics at the University of Oldenburg.

Armin Kohlrausch studied physics at the University of Göttingen, Göttingen, Germany, and specialized in acoustics and received the M.S. degree in 1980 and the Ph.D. degree in 1984, both in perceptual aspects of sound from the University of Göttingen.

From 1985 until 1990, he was with the Third Physical Institute, University of Göttingen, being responsible for research and teaching in the fields psychoacoustics and room acoustics. In 1991, he joined the Philips Research Laboratories, Eindhoven, The Netherlands, and worked in the Speech and Hearing Group of the Institute for Perception Research (IPO). Since 1998, he has combined his work at Philips Research Laboratories with a Professor position for multisensory perception at the TU/e. In 2004, he was appointed as a Research Fellow of Philips Research.

Dr. Kohlrausch is a member of a great number of scientific societies, both in Europe and the U.S. Since 1998, he has been a Fellow of the Acoustical Society of America, covering the areas of binaural and spatial hearing. His main scientific interest is in the experimental study and modelling of auditory and multisensory perception in humans and the transfer of this knowledge to industrial media applications.