Advances in Missing Feature Techniques for Robust Large-Vocabulary Continuous Speech Recognition

Maarten Van Segbroeck, *Member, IEEE*, and Hugo Van Hamme, *Member, IEEE*

***Abstract*—Missing feature theory (MFT) has demonstrated great potential for improving the noise robustness in speech recognition. MFT was mostly applied in the log-spectral domain since this is also the representation in which the masks have a simple formulation. However, with diagonally structured covariance matrices in the log-spectral domain, recognition performance can only be maintained at the cost of increasing the number of Gaussians drastically. In this paper, MFT can be applied for static and dynamic features in any feature domain that is a linear transform of log-spectra. A crucial part in MFT-systems is the computation of reliability masks from noisy data. The proposed system operates on either binary masks where hard decisions are made about the reliability of the data or on fuzzy masks which use a soft decision criterion. For real-life deployments, a compensation for convolutional noise is also required. Channel compensation in speech recognition typically involves estimating an additive shift in the log-spectral or cepstral domain. To deal with the fact that some features are considered as unreliable, a maximum-likelihood estimation technique is integrated in the back-end recognition process of the MFT system to estimate the channel. Hence, the resulting MFT-based recognizer can deal with both additive and convolutional noise and shows promising results on the Aurora4 large-vocabulary database.**

***Index Terms*—Automatic speech recognition (ASR), channel compensation, missing data techniques, noise robustness.**

I. Introduction

In contrast to human listeners, automatic speech recognition (ASR) systems are particularly sensitive to the presence of background noises and acoustical variations in the speaking environment. Speech signals processed by an ASR-system are influenced by the way they are produced (e.g., differences in speaker characteristics, speaking style, accent, dialect), by the surrounding sounds that add noise to the signal, by the transducer used to capture the signal (e.g., the microphone characteristics) and by the transmission channel (e.g., the room impulse response). In this paper, it is assumed that speaking variations and speaker variances are dealt with by the acoustic and language model of the speech recognizer.

Robustness can be defined as the ability of the ASR to maintain its performance or degrade gracefully when exposed to a range of different environmental conditions. In realistic testing situations, ASR-systems operate in environments with unknown time-varying additive noise distortions, leading to a mismatch between training and testing conditions. Considering and incorporating all possible noise mismatches, as aimed in a multi-condition training approach, is not only an impractical task but also it is often not guaranteed that the back-end model has adequate discriminative power neither is one sure to cover all relevant noise conditions. Even when the noisy environment resembles the trained conditions and is sufficiently known beforehand, it is not guaranteed that a multi-condition training approach is effective.

Hence, in the case of clean condition training, noise compensation techniques are required to reduce the model mismatch of the noisy speech by dealing with the unknown realistic noise distortions. Several approaches have been proposed in the literature which can be broadly divided into three categories. An overview of these techniques can be found in [1]–[3]. The first group of techniques aims to increase the noise robustness by extracting speech features that are inherently less distorted by the noise. Over the years, different robust feature extraction methods attempting to derive noise resistant feature parameters have been explored, e.g., perceptual linear prediction (PLP) coefficients [4], possibly combined with the relative spectra (RASTA) techniques [5], root-cepstrum coefficients (RCC) [6], the modulation spectrograms [7], spectral peaks [8] and Mel-frequency cepstral coefficients (MFCCs) [9], which later became the standard feature set in ASR. Another type of noise reduction technique operates in the feature domain by applying a transformation to the degraded speech such that the transformed speech closely resembles the clean speech. Some examples of these speech enhancement techniques are spectral subtraction [10], Wiener filtering [11], model-based enhancement techniques [12], [13], subspace-based speech enhancement methods [14], [15], and the vector Taylor series (VTS) compensation algorithm [16]. Whereas feature enhancement techniques try to remove the mismatch between the noisy observation and the acoustic model by modifying the incoming feature vector, a third category, namely the model compensation techniques modify the models learned by the recognizer such that they resemble the distribution of the observed noisy speech. Some model adaptation methods worth mentioning are the hidden Markov model (HMM) decomposition technique [17], parallel model combination (PMC) [18], maximum-likelihood linear regression (MLLR) [19], and maximum *a posterior* probability (MAP) compensation [20].

In real life situations, the statistics of the background noise are not known beforehand and are difficult to predict, while model

Manuscript received June 02, 2009; revised January 25, 2010; accepted February 01, 2010. Date of publication March 11, 2010; date of current version October 01, 2010. This work was supported in part by the Institute for the Promotion of Innovation through Science and Technology in Flanders, Belgium (I.W.T.-Vlaanderen), and in part by the MIDAS project of the Nederlandse Taalunie under the STEVIN program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nestor Becerra Yoma.

The authors are with the Department of Electrical Engineering, Katholieke Universiteit Leuven, B-3001 Leuven-Heverlee, Belgium (e-mail: maarten.vansegbroeck@esat.kuleuven.be; hugo.vanhamme@esat.kuleuven.be).

Digital Object Identifier 10.1109/TASL.2010.2045235

adaptation techniques rely on important assumptions about the noisy conditions. Due to a lack of data and their associated computational cost these techniques are rarely capable of incorporating all model parameters and thus to compensate for all noise types. Most speech enhancement techniques aim to reduce the noise by assuming that the noise is stationary over a relatively large time window. Hence, these methods mostly fail for realistic nonstationary noise distortions. In contrast, robust feature extraction makes only weak or no assumptions about the noise. However, it appears to be difficult to find a representation that is insensitive to a wide range of distortions.

It is well known that human listeners can cope with speech degradations and are capable of recognizing it by utilizing the partial information left in the distorted speech signal. From this property of humans, the *missing feature* approach originated for improving the robustness in ASR-systems. In a spectrographic representation of noisy speech, some time–frequency regions will be dominated by the noise and others are dominated by the speech. Missing feature theory (MFT) [21] attempts to compensate for additive noise distortions by first locating the corrupted time–frequency regions and then performing recognition on these partial or incomplete feature vectors. Therefore, MFT is required to estimate a missing feature mask indicating the reliability of different spectral regions in the noisy data.

MFT has already shown its effectiveness in dealing with additive noise when applied in ASR, even in challenging realistic conditions with nonstationary noise distortions. One of the reasons for the success of MFT is its similarity with the human auditory system in dealing with partly distorted data by exploiting the redundancy of the speech signal. Another advantage is that, in contrast to other noise robustness techniques, MFT does not rely on important assumptions about the noise and is intrinsically suited to compensate for either stationary or nonstationary noises. In practice however, the extent to which the performance of the MFT approach is dependent on the noise type is determined by the mask estimation procedure involved.

The principles behind MFT-based recognition will be restated in Section II. We will make a distinction between hard and soft decisions for estimating the mask. When a masking decision is hard, a binary missing data mask is produced which indicates for each time–frequency cell that it is either completely reliable or else completely missing. In the recognition process, the most commonly used approaches to deal with these partial feature vectors are marginalization [21], [22] and state-based data imputation [23], [24]. In marginalization, recognition is performed by integrating out the missing features, while in data imputation, a complete feature vector for recognition is first estimated by reconstructing the missing part from the reliable data using the HMM state distribution as *a prior*. The latter technique will be further explored in Section III since this is the technique that will be exploited in the proposed MFT-based recognizer. For a comprehensive overview of the different recognition strategies in the missing feature framework and a comparison of various aspects of these techniques, we refer to [3].

For reasons of accuracy, most ASR-systems operate in a domain that is a linear transformation of log-spectra. Therefore, the data imputation technique needs to be extended to cover such linear transformations. In [25], it has been shown that a higher accuracy and robustness can be obtained by using missing data techniques in the cepstral domain. The major drawback of this approach is that it requires significantly more computation since imputing the unreliable data requires the solution of a non-negative least squares (NNLSQ) problem. An alternative MFT formulation was presented in [26] through the introduction of the ProSpect features, such that the computational load is reduced while maintaining the accuracy. These ProSpect features will be redefined in Section III-B and the solution procedure of the NNLSQ-problem will be provided in Section IV, together with a discussion of the computational requirements in the cepstral and ProSpect domain. A comparison between the performance of the data imputation techniques in the log-spectral, cepstral and ProSpect domain is made in Section VI, based on Aurora2 digit recognition experiments.

In contrast to binary masks, soft masking decisions result in (real-valued) fuzzy mask vectors where a probability of being reliable is estimated for each spectral component. A disadvantage of the probabilistic decision framework is that the soft-bounded marginalization approach [27], [28] is the only missing feature technique that can cope with fuzzy masks, while only the imputation-based approaches allow a linear transformation of the log-spectral domain without the need of important approximations such as in [29]. Therefore, in [30], we have extended the imputation technique for binary masks to be applied in the probabilistic masking decision framework. The formulation of the optimization problem for fuzzy masks can also be found in Section III and the solution strategy will be given in Section III-B.

One of the shortcomings of MFT is that it only removes additive noise distortions. Hence, a compensation for linear filtering or convolutional noise (e.g., the mismatch between the microphone characteristics in training and testing conditions) is not present. Channel compensation in speech recognition typically involves estimating an additive shift in the log-spectral or cepstral domain and performing mean normalization. However, since some part of the data is considered as unreliable, different normalization strategies are required in MFT-based recognition. Therefore, the maximum-likelihood estimation (MLE) technique of [31] will be integrated into the MFT-based recognizer as will be explained in Section V.

This paper combines our recent findings about MFT that were proposed in previous publications. The result is a MFT-based recognizer that can deal with both additive and convolutional noise. The extended benchmarks presented in this paper show promising results. In Section VI, experimental results of the proposed MFT-based recognizer are given on the Aurora4 large vocabulary database for *a priori* and real masking strategies and the effect of the MLE channel compensation method on the word error rate is illustrated. Finally, conclusions can be found in Section VII.

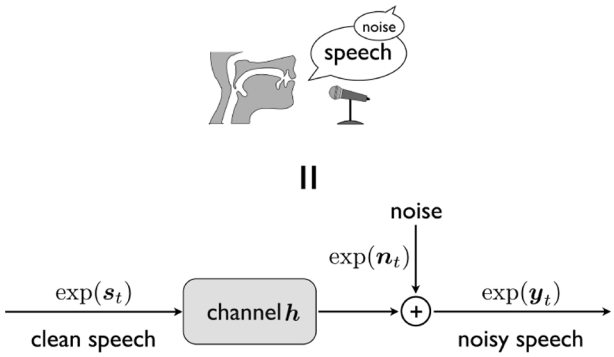

Fig. 1. Model of the acoustic environment.

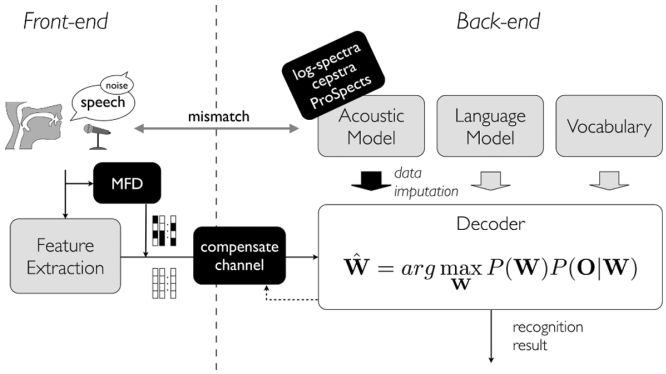

Fig. 2. Schematic representation of the proposed MFT-based recognizer.

II. Overview of Missing Feature Theory in Speech Recognition

Conventional ASR-systems consist of a front-end that performs a feature extraction and a back-end that performs the recognition of the uttered words based on the observed features by combining the information of three knowledge sources: the acoustic model, the language model and the lexicon. Let $\boldsymbol{s}_t$, $\boldsymbol{n}_t$, and $\boldsymbol{y}_t$ denote the vector of D log-Mel spectral features at a time frame t for the clean speech, the noise and the noisy signal, respectively. These feature vectors are computed by a filter bank with D channels through framing, windowing, FFT and filter bank integration. To enhance the noisy speech, an acoustic environment model is required that describes the relation between the noisy speech, the clean speech and the different sources that cause degradations. In this paper, we will deal with the two most important distortions sources: additive background noise $\boldsymbol{n}_t$ and convolutional noise denoted by the channel parameter $\boldsymbol{h}$. By assuming that the additive noise is independent of the speech, the relationship between the distorted speech $\boldsymbol{y}_t$, the additive noise $\boldsymbol{n}_t$, the channel $\boldsymbol{h}$, and the clean speech $\boldsymbol{s}_t$, is then given by

$$\boldsymbol{y}_t \approx \log(\exp(\boldsymbol{s}_t + \boldsymbol{h}) + \exp(\boldsymbol{n}_t)) \tag{1}$$

This time-domain parametric model of the acoustic environment is also shown in Fig. 1.

In conventional ASR, recognition is performed in the cepstral domain which can be obtained by applying the discrete cosine transform (DCT) matrix to the log-Mel feature vector $\boldsymbol{y}_t$. If $\boldsymbol{C}_\kappa$ denotes the κ-by-D orthonormal DCT matrix, the cepstra (or MFCCs) are defined as

$$\boldsymbol{c}_t = \boldsymbol{C}_\kappa \boldsymbol{y}_t. \tag{2}$$

In the front-end, the noisy speech is transformed into a sequence of T observation vectors $\boldsymbol{O} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_T\}$. The decoder in the back-end will then search for the word sequence hypothesis $\hat{\boldsymbol{W}}$ that maximizes the *a posteriori* probability

$$\begin{aligned}\hat{\boldsymbol{W}} &= \arg\max_{\boldsymbol{W}} P(\boldsymbol{W}|\boldsymbol{O}) \\ &= \arg\max_{\boldsymbol{W}} \frac{P(\boldsymbol{O}|\boldsymbol{W})P(\boldsymbol{W})}{P(\boldsymbol{O})}\end{aligned} \tag{3}$$

over all vocabulary word sequences $\boldsymbol{W}$ while exploiting the acoustic model $P(\boldsymbol{O}|\boldsymbol{W})$ and the language model $P(\boldsymbol{W})$. In this paper, the acoustic model is assumed to have a mainstream HMM-based architecture with Gaussian mixture models (GMMs). If N represents a multivariate Gaussian distribution, then the likelihood of the νth mixture component of HMM-state q is given by

$$f(\boldsymbol{c}_t|\nu, q) = N(\boldsymbol{c}_t|\boldsymbol{\mu}^c_{\nu,q}, \boldsymbol{\Sigma}^c_{\nu,q}) \tag{4}$$

where $\boldsymbol{\mu}^c_{\nu,q}$ and $\boldsymbol{\Sigma}^c_{\nu,q}$ are, respectively, the Gaussian mean and covariance matrix of the mixture component expressed in the cepstral domain as indicated by the superscript c. For reasons of notational convenience, we will refer to the Gaussian mixture component (ν, q) by using the subscript i. Key to the success of MFCC in ASR is that they are well modeled by a GMM with diagonal covariance due to the property that the DCT matrix attempts to decorrelate the log-spectra. Hence, the computational load during acoustic model evaluation is significantly reduced without sacrificing accuracy.

In MFT, recognition is based only on those regions in the time–frequency representation of the speech data that are matched with the recognizer's model. This requires a solution to the following problems: 1) locating of those matched regions in the time–frequency plane; 2) adaptation of the ASR-system to handle the unreliable speech information; and 3) dealing with unreliable features in convolutional noise compensation. Therefore, three important additions to the conventional architecture of the ASR are required, respectively a missing feature detector (MFD), a speech reconstruction method which exploits the recognizer's acoustic model in the back-end (e.g., a data imputation technique) and an alternative for the commonly used cepstral mean normalization method. These modifications are schematically represented in Fig. 2 and are described in the following sections. The presence of the arrow from the decoder to the channel compensation method will be explained later on.

A. Reliability Decisions

In MFT, it is assumed that reliable spectral regions can be identified with a certain probability and that they remain mostly undistorted by environmental noises. If we disregard the effect of convolutional noise, the following assumption can then be made for the noisy speech:

$$\boldsymbol{y}_t \approx \max(\boldsymbol{s}_t, \boldsymbol{n}_t) \tag{5}$$

where the max-operator works element-wise over the log-Mel spectral components. The MFD will then estimate for each spectral component of the observation $\boldsymbol{y}_t$ a probability that it is dom-

inated by speech. This way, the ASR input stream of feature vectors is augmented with spectral mask vectors indicating whether the feature component is "reliable" or "unreliable" (missing).

Decisions about the reliability of the data can be deterministic or probabilistic resulting in a different type of the mask. If a hard decision is made, the missing data mask is binary, i.e., the component is either completely reliable or else completely missing. *Ideal* binary masks are obtained by comparing the log-Mel spectra of clean speech and noise

$$\boldsymbol{m}_t = (\boldsymbol{s}_t \geq \boldsymbol{n}_t + \theta)_{0/1} \tag{6}$$

where $(\ldots)_{0/1}$ equals 1 (0) when the logical expression inside the brackets holds (does not hold) and θ is a constant threshold. If 1 is assigned to a time–frequency cell, it is dominated by speech, while the mask value 0 indicates that the cell is masked by background noise. In realistic situations however, $\boldsymbol{s}_t$ and $\boldsymbol{n}_t$ are unknown and the mask has to be estimated. The accuracy of real binary masks is crucial since mask estimation errors will cause a significant degradation in recognition performance.

The alternative is to use a soft decision criterion by estimating for each spectral component a probability of being dominated by the speech. The obtained mask is then called soft or fuzzy and the fuzzy mask vector $\boldsymbol{w}_t$ can be generated by the approach of [27], i.e., by the substitution of (6) in a sigmoid function

$$\boldsymbol{w}_t = \left(\frac{1}{1+\exp(-\rho(\boldsymbol{s}_t - \boldsymbol{n}_t - \theta))}\right) \tag{7}$$

with slope ρ. The mask vector of (7) now consists of continuous values between 0 and 1. If the value is close to 1, the component has a high probability of being dominated by speech.

Previous research ([32], [27], and [33]) has shown that soft masking decisions give better results than hard decisions. Another argument in favor of fuzzy masks is that it is not guaranteed that binary masks identify a sufficient quantity of partial information for recognition of the speech. However, fuzzy masks also have some disadvantages. Experiments have shown that the parameter ρ in (7), which expresses the slope of the sigmoid function, is often highly sensitive to the noise type. Another drawback is that if the data imputation technique in the back-end is applied in a domain that is a linear transform of the log-Mel spectral domain, the solution strategy becomes computationally more complex. The discussion about the mask type to be used will be continued in Section VI by interpreting the final recognition results.

In most ASR-systems, the static feature vector $\boldsymbol{y}_t$ is augmented by its first- and second-order derivatives, i.e., a velocity and acceleration feature vector. We have already shown in [34] that MFT can also be applied on the dynamic features. In contrast to statics, dynamic spectra that are corrupted by noise can result in a noisy observation value that is either larger or smaller than the clean value. Hence, the MFD will need to generate a ternary mask where

$$\Delta\boldsymbol{m}_t = (\partial\boldsymbol{s}_t \geq \partial\boldsymbol{y}_t + \theta)_{0/1} + 2(\partial\boldsymbol{s}_t \leq \partial\boldsymbol{y}_t - \theta)_{0/1} \tag{8}$$

where ∂ denotes either the first or second order derivative operator. A graphical illustration of (8) is given in Fig. 3. Dynamic

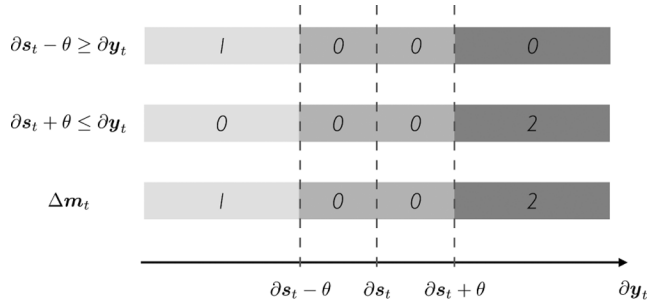

Fig. 3. Graphical illustration of the ternary decision for obtaining ideal delta masks.

components are considered reliable if they deviate less than θ from the clean values. These components have a mask value 0. Unreliable dynamic feature components have mask values 1 or 2 and are, respectively, imputed with the constraints $\partial\boldsymbol{s}_t \leq \partial\boldsymbol{y}_t$ or $\partial\boldsymbol{s}_t \geq \partial\boldsymbol{y}_t$. This approach was successfully tested in [34] using *a priori* masks derived from knowledge of the clean speech and the noise. Here, it was also illustrated that (suboptimal) real dynamic masks can be obtained from an ad hoc derivation of the binary static mask; hence,

$$\Delta\boldsymbol{m}_t = \begin{cases} 0, & \text{if } \partial\boldsymbol{m}_t = 0 \\ 1, & \text{if } \partial\boldsymbol{m}_t > 0 \\ 2, & \text{if } \partial\boldsymbol{m}_t < 0. \end{cases} \tag{9}$$

The above equation can be interpreted as follows. The dynamic spectra are a linear combination of the static spectra. The static masks flag the fact that the noisy contribution to this linear combination is less than the clean value. This noise contribution will cause the value of the dynamic features to be higher $(\Delta\boldsymbol{m}_t = 1)$ or lower $(\Delta\boldsymbol{m}_t = 2)$ than the clean value. If there are equal votes for over and underestimation, i.e., when all features are reliable or unreliable due to the noise corruption, the dynamic features are considered as reliable and $\Delta\boldsymbol{m}_t = 0$.

Examples of a binary and fuzzy static mask, together with the masks for the dynamic features are shown in Fig. 4 for a noisy utterance of the Aurora4 database mixed with airport noise at 10-dB SNR. The noisy time–frequency representation is perceptually smoothed by a 22-channel Mel-scale filter bank, computed every 10 ms using a frame window length of 32 ms. The static masks of Fig. 4 are *a priori* masks. In the remainder of the paper, we will mainly focus on the back-end of the MFT-based recognizer. In Section VI, its performance will be evaluated in terms of word error rate obtained with the real masks that were proposed in [35]. Here, it was attempted to exploit as much *a priori* speech knowledge as possible while restricting the noise modeling to a minimum. A vector quantization (VQ) strategy exploiting the harmonicity structure of speech was used. The noise is assumed to contain only weak harmonicity structures and is estimated using a harmonic tunneling approach [36]. The noisy speech is segregated into a clean speech and noise stream, which are subsequently compared to form a mask. To compensate for linear channel distortions, the VQ-system will self-adjust the codebook containing the codewords to the channel during online recognition. For an overview of other mask estimation methods that were described in literature, we refer to [37].

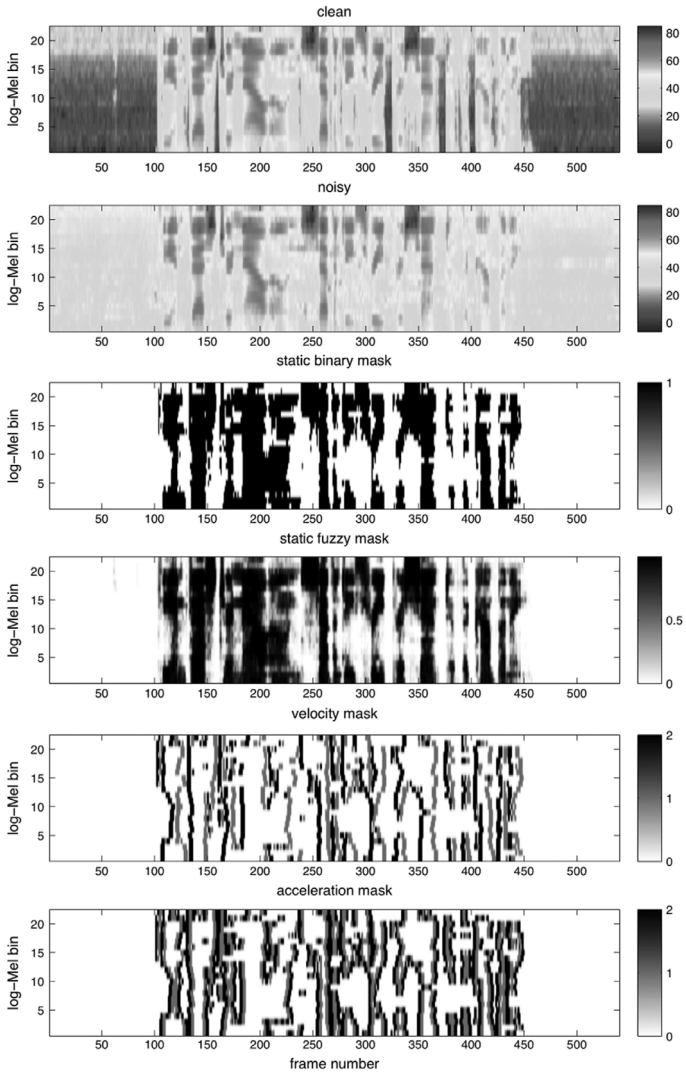

Fig. 4. From top to bottom: the log-Mel spectrogram for the clean and noisy utterance "Analysts haven't focused on what happened to them." of the Aurora4 database corrupted by airport noise at 10-dB SNR. The binary mask for the static features created using *a priori* knowledge of the speech and the noise in (6) with an SNR threshold θ of -3 dB. The corresponding fuzzy mask for the static features obtained by (7) using a slope value ρ of 0.1. The mask for the velocity features and for the acceleration features computed by the first and second-order derivative of the binary mask.

B. Compensation for Unreliable Data

A second modification to the ASR-system needs to be made during the evaluation of the acoustic model in the back-end using MFT. Here, the probability of being reliable of each feature component should be taken into account while computing the acoustic scores. A major drawback of MFT is that the acoustic model must be expressed in the spectral domain since this is also the representation in which the masks have a simple formulation. While the cepstral domain is often used for reasons of accuracy, conventional MFT-techniques like bounded marginalization [21], [22], or the imputation techniques of [23], [24], rely on GMMs with diagonal covariance expressed in the log-spectral domain. Therefore, a data imputation technique was introduced in [25] where the spectral representation can be replaced by any linear transform of the log-spectra.

By using MFT-techniques with cepstra, a superior accuracy and robustness relative to their spectral competitors is obtained. The price to be paid is that the imputation of the missing data is more complex: the evaluation of a Gaussian now requires the solution of a non-negative least squares (NNLSQ) problem. Through the introduction of the ProSpect features [26], the computational load of the cepstral representation is alleviated while their accuracy is maintained. These MFT-techniques were introduced to be used with binary missing data masks and were extended in [30] such that they can cope with fuzzy masks.

In Section III, the maximum-likelihood-based imputation technique per Gaussian is restated. Here, the optimization function will be formulated such that it can be solved in the log-Mel spectral domain or in any other domain that is a linear transformation of log-Mel spectral features. The procedure for optimizing the cost function using ProSpect and cepstral features is given in Section III-B. Here, a distinction will be made between the solution strategy for binary and fuzzy masks. The approach for compensating delta features in the MFT framework will be discussed as well.

C. Dealing With Convolutional Noise

Missing Data Techniques were originally formulated as a technique to compensate for additive noise distortions. For real-life deployments, a compensation for unknown filtering is also required. Convolutional noise is caused by variations in the transmission channel, such as the use of a different microphone than the one used during training, changes in the distance between the microphone and the speaker or modifications in the recording environment. Most of these variations can be modeled as a linear filtering, i.e., convolutional noise, and as a translation of the log-spectral features. In this paper, channel variations are limited to short impulse responses and room reverberation is not considered.

From the acoustic environment model (1) it is clear that a GMM for $\boldsymbol{s}_t$ that was trained on undistorted data, can be matched to the distorted data by adding a shift $\boldsymbol{h}$ to the clean speech means. The conventional strategy in ASR to compensate for convolutional noise is to subtract the cepstral mean from the observed speech data, i.e., cepstral mean subtraction (CMS). Since cepstral and log-Mel spectral features are also related to each other by a linear transform, this operation can also be done in the log-Mel spectral domain (log-Mel MS). A disadvantage of mean subtraction methods is when some log-Mel spectral features are not attributed to speech, but to a different source, as is done in MFT, simple averaging will create an important bias and mean subtraction methods loose their effect. This bias arises from the fact that the additive noise source will also contribute to the mean value. This is undesirable for the MFT-framework since it requires that the normalized reliable speech regions match the clean speech on which the acoustic model was trained. This problem can be solved in the MFT framework by taking the information of the missing feature mask into account. This way, the components that are distorted by the noise can be left out of consideration in the computation of the channel. Conventional techniques suffer from bias due to noise. MFT offers—at least for oracle masks—to have a bias-free estimation. As the examples of the Section VI will

show, it can even handle narrowband data, i.e., when some frequency regions are always missing. Conventional techniques result in erroneous results in this case.

A method that is compatible with missing data has been proposed in [38], in which the spectral features are normalized by a factor computed only from the most intense regions of the speech. However, when the data becomes band limited, such as during conference interviews or other audio mining applications, no speech information will be available in certain frequency regions and then the normalization method of [38] will become inappropriate. Therefore, we have implemented a technique in our MFT-based recognition system that exploits the back-end speech model to estimate the channel by means of an MLE algorithm [31]. Here, the initial channel estimate is updated by maximizing the log-likelihood of the optimal state sequence of the observation data.

The use of the back-end speech model, trained on clean speech data, results in two main advantages compared to the method of [38]: 1) the channel compensation method can handle narrowband data and 2) it produces a channel estimate with a negligible bias arising from the noise. An outline of this method will be restated in Section V.

III. Acoustic Model Evaluation in MFT

In this section, the data imputation strategy will be presented as a Gaussian likelihood minimization problem such that it can be applied in any feature domain that is a linear transform of log-spectra. The optimization function is then investigated in further detail in case of the log-spectral and cepstral domain. ProSpect features will also be defined and their computational advantages will be made clear in Section IV.

A. Data Imputation

For each time frame t, the maximum-likelihood based imputation technique computes an estimate for the speech from the information contained in the corresponding observation and the mask vector, while using the acoustic back-end model as a prior. Hence, the imputation technique produces a Gaussian-dependent estimate for the speech. Afterwards, the acoustic scores are computed by evaluating the Gaussians in the obtained speech estimates.

If the mask is binary, the reliable components $\boldsymbol{s}_{t,r}$ of the clean speech are approximated by their counterparts in the noisy speech $\boldsymbol{y}_{t,r}$, while the unreliable speech components $\boldsymbol{s}_{t,u}$ are considered as unknown. Hence, the missing part of $\boldsymbol{s}_t$ will be estimated by minimizing the negative log-likelihood Φ for each Gaussian mixture component i over $\boldsymbol{s}_t$ [25]

$$\begin{aligned} \Phi_{t,i} &= \frac{1}{2}(\boldsymbol{s}_t - \boldsymbol{\mu}_i)' \boldsymbol{P}_i (\boldsymbol{s}_t - \boldsymbol{\mu}_i) \\ \text{s.t. } \boldsymbol{s}_{t,r} &= \boldsymbol{y}_{t,r} \text{ and } \boldsymbol{s}_{t,u} \leq \boldsymbol{y}_{t,u}. \end{aligned} \tag{10}$$

Here, $\boldsymbol{P}_i$ is an inverse covariance or precision matrix of a Gaussian that is expressed in the log-spectral domain or in any other domain that is a linear transformation of log-spectral features, e.g., cepstra or ProSpect features (see Section III-B). Despite their performance differences, all these variants of MFT have a known symmetric positive-definite precision matrix. In (10), the mean of the Gaussian transformed into the log-Mel domain is given by $\boldsymbol{\mu}_i$. Both parameters $\boldsymbol{P}_i$ and $\boldsymbol{\mu}_i$ are estimated on clean training data.

As mentioned in Section II-A, the type of the mask can be fuzzy and dealing with this different source of reliability information about the speech data will result in a different likelihood function that has to be optimized during data imputation [30]. In a fuzzy masking approach, the data is not partitioned into a reliable and unreliable part as is the case with binary masks. Moreover, the soft reliability information has to be taken into account in the optimization function. Assuming the use of Gaussians with a diagonal covariance matrix, this function should have the property that 1) if the mask value is close to 1, the optimal point tends to the observation value, and 2) if the mask value is close to 0, the optimal point tends to a value as close to the Gaussian mean as permitted by the constraint $\boldsymbol{s}_t \leq \boldsymbol{y}_t$. Condition 2) is fulfilled if the precision matrix $\boldsymbol{P}_i$ in (10) is replaced by

$$\boldsymbol{Q}_{t,i} = (\boldsymbol{I}_D - \boldsymbol{W}_t)^{\frac{1}{2}} \boldsymbol{P}_i (\boldsymbol{I}_D - \boldsymbol{W}_t)^{\frac{1}{2}} \tag{11}$$

where $\boldsymbol{I}_D$ denotes the $D \times D$ identity matrix and $\boldsymbol{W}_t$ is the $D \times D$ diagonal matrix with the elements of $\boldsymbol{w}_t$ of (7) on the diagonal. Note that the matrix multiplications in (11) preserve the symmetry of $\boldsymbol{P}_i$. This modification together with the addition of the extra term $(1)/(2)(\boldsymbol{s}_t - \boldsymbol{y}_t)' \boldsymbol{W}_t (\boldsymbol{s}_t - \boldsymbol{y}_t)$, will meet condition 1). The optimization problem for fuzzy masks then becomes

$$\begin{aligned} \Psi_{t,i} &= \frac{1}{2}(\boldsymbol{s}_t - \boldsymbol{\mu}_i)' \boldsymbol{Q}_{t,i} (\boldsymbol{s}_t - \boldsymbol{\mu}_i) \\ &\quad + \frac{1}{2}(\boldsymbol{s}_t - \boldsymbol{y}_t)' \boldsymbol{W}_t (\boldsymbol{s}_t - \boldsymbol{y}_t) \\ \text{s.t. } \boldsymbol{s}_t &\leq \boldsymbol{y}_t \end{aligned} \tag{12}$$

Note that all constraints are now inequality constraints, since there is no evidence for a specific component of being reliable.

B. MFT in Different Feature Domains

The optimization of functions (10) and (12) has to be performed for all different Gaussians of the acoustic model to produce a Gaussian-dependent estimate $\hat{\boldsymbol{s}}_{t,i}$ for the clean speech. Hence, the computational complexity of the optimization problem is an important concern in the MFT framework. The MFT variants in the spectral, cepstral and ProSpect domain are discussed next.

1) Log-Mel Spectra: In the spectral domain, the precision matrix $\boldsymbol{P}_i$ of Gaussian i is the inverse of the log-Mel spectral covariance matrix $\boldsymbol{\Sigma}_i^s$ of size $D \times D$. The superscript s will be used to indicate that the covariance matrix is expressed in the spectral domain. Hence, (10) and (12) have a simple solution since they can be decomposed in D independent problems. If the mask is binary, each jth component of $\hat{\boldsymbol{s}}_{t,i}$ at time t for mixture component i is then computed as

$$\hat{s}_{t,i,j} = \begin{cases} (1 - m_{t,j})\mu_{i,j} + m_{t,j} y_{t,j}, & \text{if } y_{t,j} > \mu_{i,j}, \\ y_{t,j}, & \text{otherwise.} \end{cases} \tag{13}$$

The components of the optimal point $\hat{\boldsymbol{s}}_{t,i}$ that minimizes the fuzzy mask optimization function (12) are given by

$$\hat{s}_{t,i,j} = \begin{cases} \bar{\mu}_{t,i,j}, & \text{if } y_{t,j} > \bar{\mu}_{t,i,j}, \\ y_{t,j}, & \text{otherwise} \end{cases} \tag{14}$$

where

$$\bar{\mu}_{t,i,j} = \frac{(1-w_{t,j})\mu_{i,j}/\sigma^2_{i,j} + w_{t,j}y_{t,j}}{(1-w_{t,j})/\sigma^2_{i,j} + w_{t,j}} \tag{15}$$

and with $\sigma^2_{i,j}$ the jth diagonal element of $\boldsymbol{\Sigma}^s_i$.

2) Cepstra: The loss in accuracy due to the diagonal covariance matrix of the spectral features can be overcome by expressing the MFT problem in the cepstral domain. The precision matrix of Gaussian i is here defined as [26]

$$\boldsymbol{P}_i = \boldsymbol{C}'_\kappa \left[\boldsymbol{\Sigma}^c_i\right]^{-1} \boldsymbol{C}_\kappa + \epsilon \boldsymbol{\Sigma}^s_i \tag{16}$$

the second term is needed for regularization and ϵ is a noncritical constant. Without the regularization term, the precision matrix would be rank-deficient ($\kappa \leq D$) which causes the optimization problem to be undetermined.

The optimization of (10) and (12) has now become computationally more complex since they are formulated as a NNLSQ-problem, i.e., the constrained minimization of a quadratic. An iterative solution strategy will be presented in Section IV.

3) ProSpects: ProSpect features are an alternative representation to cepstral features and were presented in [26]. Just like cepstra, they are computed by a linear transform of the logarithm of the filter bank energies. While they can be applied in any speech recognition system, they show especially a clear benefit in MFT-based recognition since they reduce the computational requirements over CMFT while the accuracy is maintained. This will be illustrated in Section IV-B by comparing the computational complexity of solving problems (10) or (12) in the cepstra and ProSpect domain.

If we consider a λ-dimensional cepstral feature vector $\boldsymbol{c}_t = \mathbf{C}_\lambda \boldsymbol{y}_t$, the residual spectrum $\boldsymbol{d}_t$ is

$$\boldsymbol{d}_t = \boldsymbol{y}_t - \boldsymbol{C}'_\lambda \boldsymbol{c}_t = (\boldsymbol{I}_D - \boldsymbol{C}'_\lambda \boldsymbol{C}_\lambda)\boldsymbol{y}_t \tag{17}$$

where $\boldsymbol{C}'_\lambda \boldsymbol{C}_\lambda$ is a projection matrix since $\boldsymbol{C}_\lambda$ is orthonormal. Hence, $\boldsymbol{d}_t$ is the projection of $\boldsymbol{y}_t$ onto the space perpendicular to the space spanned by the rows of $\boldsymbol{C}_\lambda$. The ProSpect features are now defined as

$$\boldsymbol{p}_t = \begin{bmatrix} \boldsymbol{c}_t \\ \boldsymbol{d}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_\lambda \\ \boldsymbol{I}_D - \boldsymbol{C}'_\lambda \boldsymbol{C}_\lambda \end{bmatrix} \boldsymbol{y}_t = \begin{bmatrix} \boldsymbol{C}_\lambda \\ \boldsymbol{D}^\perp_\lambda \end{bmatrix} \boldsymbol{y}_t. \tag{18}$$

In [26], it has been shown that these features can be modeled well by a GMM with diagonal covariance, even for λ as small as 3. The likelihood of the νth mixture component of HMM state q has the expression

$$f(\boldsymbol{p}_t|\nu,q) = N\left(\boldsymbol{c}_t|\boldsymbol{\mu}^c_{\nu,q}, \boldsymbol{\Sigma}^c_{\nu,q}\right) \cdot N\left(\boldsymbol{d}_t|\boldsymbol{\mu}^d_{\nu,q}, \boldsymbol{\Sigma}^d_{\nu,q}\right)^\gamma \tag{19}$$

where γ is a stream exponent and the superscripts c, d, and p are, respectively, used to denote cepstral, residual or ProSpect features. The means and covariances in a ProSpect model are estimated using the EM algorithm where $\boldsymbol{\Sigma}^c_{\nu,q}$ and $\boldsymbol{\Sigma}^d_{\nu,q}$ are diagonal and of size $\lambda \times \lambda$ and $D \times D$, respectively.

Again using the subscript i to denote mixture component (ν, q), the precision matrix for the Gaussian of (19) is then given as [26]

$$\boldsymbol{P}_i = \boldsymbol{C}'_\lambda \left[\boldsymbol{\Sigma}^c_i\right]^{-1} \boldsymbol{C}_\lambda + \gamma \boldsymbol{D}^\perp_\lambda \left[\boldsymbol{\Sigma}^d_i\right]^{-1} \boldsymbol{D}^\perp_\lambda \tag{20}$$

Hence, the ProSpect model defines a particular structure of the precision matrix of the spectral features, containing only $\lambda + D$ variance parameters (γ is fixed).

IV. Solving the NNLSQ-Problem

Since the precision matrices are non-diagonal in the cepstral and ProSpect domain, the quadratic cost functions (10) and (12) are associated with an equivalent non-negative least square problem (NNLSQ), i.e., the minimization over vector $\boldsymbol{x}$ of

$$\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u})'\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{u}) + \boldsymbol{b}'(\boldsymbol{x}-\boldsymbol{u}) + c \text{ s.t. } \boldsymbol{x} \geq 0 \tag{21}$$

where $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{u}$ and c are given Gaussian-dependent constants. In this section, we describe the procedure for solving the NNLSQ-problem using the gradient descent method for the optimization problems using binary static, fuzzy static and ternary dynamic masks. Next, the involved computational requirements in the cepstral and ProSpect domain will be compared.

A. Optimization Strategy

1) Binary Static Masks: To avoid costly matrix inversions, the NNLSQ-problem will be solved by the gradient descent method discussed in [26]. Therefore, the search is started from the spectral MFT solution (13). In each iteration k, the initial values are updated by

$$\boldsymbol{s}^{(k+1)}_{t,i} = \boldsymbol{s}^{(k)}_{t,i} - \hat{\alpha} \nabla \boldsymbol{\phi}^{(k)}_{t,i} \tag{22}$$

where the step direction $\nabla \boldsymbol{\phi}^{(k)}_{t,i}$ is derived from the cost gradient $\nabla \Phi^{(k)}_i = \boldsymbol{P}_i(\boldsymbol{s}^{(k)}_{t,i} - \boldsymbol{\mu}_i)$ by zeroing out those components that 1) are labeled as reliable or 2) where the gradient is negative and the corresponding speech estimate is on the constraint boundary. The optimal step size is given by

$$\alpha = \frac{\nabla \boldsymbol{\phi}^{(k)\prime}_{t,i} \nabla \boldsymbol{\phi}^{(k)}_{t,i}}{\nabla \boldsymbol{\phi}^{(k)\prime}_{t,i} \boldsymbol{P}_i \nabla \boldsymbol{\phi}^{(k)}_{t,i}} \tag{23}$$

and is reduced to $\hat{\alpha}$ such that all components of $\boldsymbol{s}^{(k+1)}_{t,i}$ satisfy the constraint $\boldsymbol{s}^{(k+1)}_{t,i} \leq \boldsymbol{y}_t$; hence,

$$\hat{\alpha} = \min\left[\min\left(\alpha \nabla \boldsymbol{\phi}^{(k)}_{t,i}, \boldsymbol{y}_t - \boldsymbol{s}^{(k)}_{t,i}\right) \cdot / \nabla \boldsymbol{\phi}^{(k)}_{t,i}\right] \tag{24}$$

where $\cdot/$ denotes the element-wise division. Since (10) is a convex problem, the gradient descent method will always find the optimal solution when enough iteration steps are applied. Experiments have shown that the gradient descent method is sufficiently converged in terms of recognition accuracy in 1 or

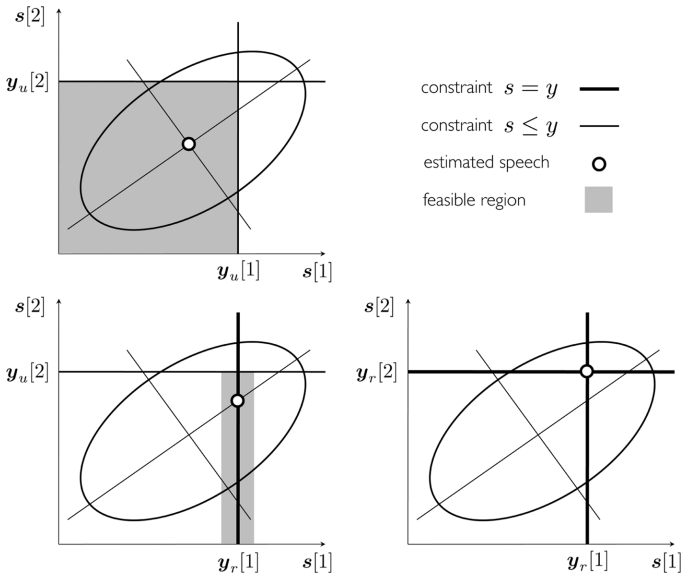

Fig. 5. Graphical illustrations of the data imputation technique for binary masks in case of a two-dimensional feature vector.

2 $(= K)$ iterations [26]. More iterations could be required if different initial values were chosen.

Finally, we obtain a clean speech estimate $\hat{\boldsymbol{s}}_{t,i} = \boldsymbol{s}_{t,i}^{(K)}$ for each Gaussian i, from which we can compute the corresponding likelihood

$$f(\hat{\boldsymbol{s}}_{t,i}|i) = \sqrt{\frac{|\boldsymbol{P}_i|}{2\pi^D}} e^{-\frac{1}{2}(\hat{\boldsymbol{s}}_{t,i}-\boldsymbol{\mu}_i)'\boldsymbol{P}_i(\hat{\boldsymbol{s}}_{t,i}-\boldsymbol{\mu}_i)}. \quad (25)$$

A graphical illustration of the constrained minimization of (10) is depicted in Fig. 5 in case of a two-dimensional feature vector. Note that the shown Gaussians have a full covariance matrix: they correspond to Gaussians in, e.g., the cepstral domain which are transformed back to the log-spectral domain. The estimated speech, the solution of the constrained optimization of (10), is indicated for three different situations in which the constraints can be active ($\boldsymbol{s}_r = \boldsymbol{y}_r$ or $\boldsymbol{s}_u = \boldsymbol{y}_u$) or passive ($\boldsymbol{s}_u < \boldsymbol{y}_u$). If the two components are labeled as unreliable, both constraints are passive (top panel) and the speech is estimated as the Gaussian mean since it lies in the feasible region and thus minimizes (10). In case the second component is reliable, the corresponding constraint is active (bottom left panel) and the optimal point has to lay on that boundary. Finally, if both components are labeled as reliable (bottom right panel), they are approximated by their corresponding observation values.

2) Fuzzy Static Masks: The optimization problem (12) is also a constrained minimization of a quadratic; hence, a gradient descent method similar to the one explained above is used for solving the problem in the cepstral or ProSpect domain. A good choice to initialize the search is to start from the log-Mel spectral solution (14) or from the point

$$\hat{\boldsymbol{s}}_{t,i}^{(0)} = \min(\boldsymbol{\mu}_i + \boldsymbol{w}_t(\boldsymbol{y}_t - \boldsymbol{\mu}_i), \boldsymbol{y}_t). \quad (26)$$

This starting point is then iteratively updated by the rule

$$\boldsymbol{s}_{t,i}^{(k+1)} = \boldsymbol{s}_{t,i}^{(k)} - \hat{\beta}\nabla\boldsymbol{\psi}_{t,i}^{(k)}. \quad (27)$$

The step direction $\nabla\boldsymbol{\psi}_{t,i}^{(k)}$ is equal to the cost gradient

$$\nabla\Psi_{t,i}^{(k)} = \boldsymbol{Q}_{t,i}\left(\boldsymbol{s}_{t,i}^{(k)} - \boldsymbol{\mu}_i\right) + \boldsymbol{W}_t\left(\boldsymbol{s}_{t,i}^{(k)} - \boldsymbol{y}_t\right) \quad (28)$$

for those components where the corresponding speech estimate lies below the constraint boundary, otherwise the component of $\nabla\boldsymbol{\psi}_{t,i}^{(k)}$ will be set to zero. The optimal step size is given by

$$\beta = \frac{\nabla\boldsymbol{\psi}_{t,i}^{(k)\prime}\nabla\boldsymbol{\psi}_{t,i}^{(k)}}{\nabla\boldsymbol{\psi}_{t,i}^{(k)\prime}(\boldsymbol{Q}_{t,i} + \boldsymbol{W}_t)\nabla\boldsymbol{\psi}_{t,i}^{(k)}} \quad (29)$$

and is reduced to $\hat{\beta}$ such that all components of $\boldsymbol{s}_{t,i}^{(k+1)}$ satisfy the constraint $\boldsymbol{s}_{t,i}^{(k+1)} \leq \boldsymbol{y}_t$; hence,

$$\hat{\beta} = \min\left[\min\left(\beta\nabla\boldsymbol{\psi}_{t,i}^{(k)}, \boldsymbol{y}_t - \boldsymbol{s}_{t,i}^{(k)}\right) \cdot / \nabla\boldsymbol{\psi}_{t,i}^{(k)}\right]. \quad (30)$$

Convergence is again reached after 1 or 2 iteration steps. Finally, the likelihood of each Gaussian i of the acoustic model is obtained by the substitution of $\hat{\boldsymbol{s}}_{t,i}$ in (25).

3) Ternary Dynamic Masks: The decorrelation properties of the ProSpect representation hold equally well for the dynamic features; hence, they are also modeled using a GMM with diagonal covariance matrices. As for binary static masks, the dynamic feature vector is divided into a reliable $\partial\boldsymbol{s}_{t,r}$ and unreliable part $\partial\boldsymbol{s}_{t,u}$. The maximum-likelihood estimate for the unreliable dynamic feature components are then found by minimizing a cost function of form (10) but now subject to the constraints

$$\begin{cases} \partial\boldsymbol{s}_{t,r} = \partial\boldsymbol{y}_{t,r}, & \text{if } \Delta\boldsymbol{m}_t = 0 \\ \partial\boldsymbol{s}_{t,u} \leq \partial\boldsymbol{y}_{t,u} + \theta, & \text{if } \Delta\boldsymbol{m}_t = 1 \\ \partial\boldsymbol{s}_{t,u} \geq \partial\boldsymbol{y}_{t,u} - \theta, & \text{if } \Delta\boldsymbol{m}_t = 2. \end{cases} \quad (31)$$

Similarly as in Section IV-A1, the step direction is derived from the gradient of the cost function by zeroing out those components that 1) are reliable or 2) are unreliable with $\Delta\boldsymbol{m}_t = 1(\Delta\boldsymbol{m}_t = 2)$ and where the gradient is negative (positive) and the corresponding speech estimate is on the constraint boundary $\partial\boldsymbol{s}_t \simeq \partial\boldsymbol{y}_t$.

B. Computational Complexity

The computational difference of solving the NNLSQ-problem in the cepstral or ProSpect domain is due to the matrix multiplication with the precision matrix (16) or (20) that is involved in each gradient step. The multiplication is required in the computation of the gradient and of the step size and can be computed from the items listed in Table I. Here, the vector $\boldsymbol{z}$ represents either $(\boldsymbol{s}_t^{(k)} - \boldsymbol{\mu}_i)$ or $\nabla\boldsymbol{\psi}^{(k)}$ and U is defined as the number of feature components that are labeled as missing in the binary masks $(U \leq D)$. The computational requirements are shown in terms of multiply-accumulate operations while neglecting $\mathrm{O}(D)$, $\mathrm{O}(U)$, and $\mathrm{O}(\lambda)$ operations, e.g., multiplying with a diagonal matrix. The resulting cost in a practical situation with $D = 23, U = 16$ (averaged measured on the data) and $\lambda = 13$ in case of cepstral features or $\lambda = 3$ for ProSpects, is given in the last row of the table. Note that in a fuzzy masking strategy, all features are taken into account during optimization and thus $U = D$. These results show that

TABLE I
APPROXIMATION OF THE COMPUTATIONAL COST OF A SINGLE GRADIENT ITERATION FOR THE BINARY AND FUZZY MASK APPROACH IN THE CEPSTRAL ($\lambda = 13$) AND PROSPECT DOMAIN ($\lambda = 3$)

Operation	Binary masks		Fuzzy masks	
	cepstra	ProSpects	cepstra	ProSpects
$\boldsymbol{C}_\lambda \boldsymbol{z}$	λU	λU	λD	λD
$\boldsymbol{C}'_\lambda [\boldsymbol{\Sigma}^c]^{-1} \boldsymbol{C}_\lambda \boldsymbol{z}$	λU	λU	λD	λD
$(\boldsymbol{I}_D - \boldsymbol{C}'_\lambda \boldsymbol{C}_\lambda)\boldsymbol{z}$	-	λD	-	λD
$\boldsymbol{D}^\perp_\lambda [\boldsymbol{\Sigma}^d_i]^{-1} \boldsymbol{D}^\perp_\lambda \boldsymbol{z}$	-	$\lambda(U+D)$	-	$2\lambda D$
Total per iteration	$4\lambda U$	$2\lambda(3U+2D)$	$4\lambda D$	$10\lambda D$
in practice	832	564	1196	690

the computational cost can be reduced with approximately 30%–40% by exploiting the structure of the precision matrix in the ProSpect domain. The computational requirements for the dynamic features using ternary masks are equivalent to those of the binary static mask optimization problem.

V. CONVOLUTIONAL NOISE COMPENSATION

During the decoding process of the recognizer, the channel parameters are estimated by maximizing the log-likelihood of the optimal state sequence

$$\hat{Q} = \arg \max_Q P(Q|\boldsymbol{Y}) = \{q_1, q_2, \ldots, q_T\} \tag{32}$$

given T successive observation vectors $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$. A proper choice for this T will be discussed later on. The log-likelihood of $\hat{Q}$ can be approximated by considering for every state q_t only the most dominant Gaussian of the mixture, i.e., the Gaussian with highest likelihood. This way, only one Gaussian at each time t has to be taken into account and the corresponding Gaussian mean and precision matrix will be denoted by, respectively, $\boldsymbol{\mu}_t$ and $\boldsymbol{P}_t$. The MLE of the channel $\boldsymbol{h}$ can then be obtained by minimizing the cost function $L(\boldsymbol{h})$ over $\boldsymbol{h}$

$$L(\boldsymbol{h}) = \sum_{t=1}^{T} \frac{1}{2} (\hat{\boldsymbol{s}}_t - \boldsymbol{\mu}_t - \boldsymbol{h})' \boldsymbol{P}_t (\hat{\boldsymbol{s}}_t - \boldsymbol{\mu}_t - \boldsymbol{h}) \tag{33}$$

while evaluating each Gaussian of (33) in its optimal point $\hat{\boldsymbol{s}}_t$, namely the static clean speech estimate. This estimate was obtained by the MFT technique as described in Section IV while using the reliability information provided by the mask. This way, the missing feature mask also has an impact on the channel estimate by the influence it has on $\hat{\boldsymbol{s}}_t$. Note that $\hat{\boldsymbol{s}}_t$ is a function of $\boldsymbol{h}$ and that $L(\boldsymbol{h})$ depends on the sequence of dominant Gaussians. Hence, iterative optimization is required. Using the Newton–Raphson method, the estimate for the channel shift $\Delta\boldsymbol{h}$ can be found as

$$\Delta\boldsymbol{h} = (\nabla^2 L(\boldsymbol{h}))^{-1} \nabla L(\boldsymbol{h}) \tag{34}$$

where $\nabla L(\boldsymbol{h})$ and $\nabla^2 L(\boldsymbol{h})$, respectively, denotes the gradient and Hessian of $L(\boldsymbol{h})$. This channel update should be applied recursively until convergence, which would imply several recognition passes for a given utterance. Fortunately, experiments have shown that one iteration per T frames suffices to guarantee convergence when the initial channel estimate is chosen as the mean over the first 200 log-Mel speech frames. A derivation of the gradient and Hessian of (34) was outlined in [31] and is expanded in Appendix I. Once $\Delta\boldsymbol{h}$ is obtained, we can compensate for it during further evaluation of the acoustic model by incorporating the channel estimate in the Gaussian-based imputation functions (10) and (12). This is equivalent to shifting the mean of all Gaussians i by $\Delta\boldsymbol{h}$. A diagram of the convolutional channel compensation procedure is shown in Fig. 6.

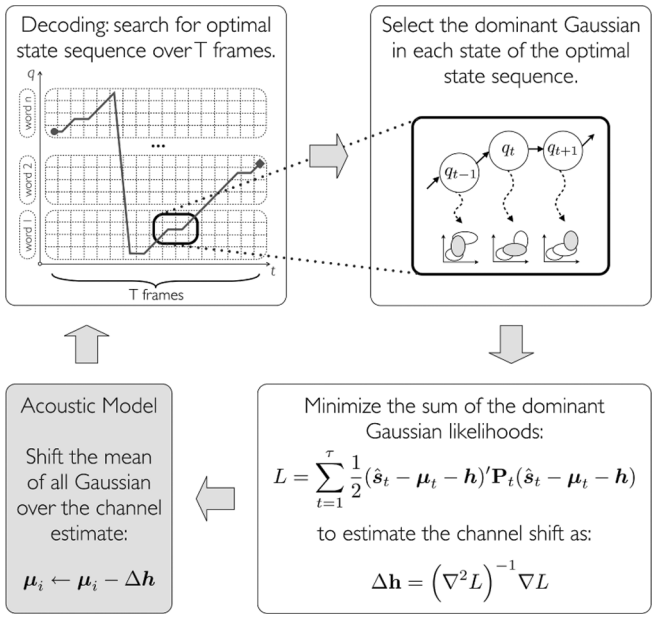

Fig. 6. Diagram of the maximum-likelihood-based channel compensation method used in the MFT-based recognizer.

Since only one Gaussian has to be evaluated per frame t, the computational load to recompute $\hat{\boldsymbol{s}}_t$ is negligible compared to the total time spent on the evaluation of all Gaussians during recognition. Furthermore, as explained in Appendix I, (34) can be efficiently computed by exploiting a QR-decomposition.

The length T of the observation sequence is dynamically chosen to ensure that we have collected a sufficient amount of speech data such that the selected Gaussians are representative for a variety of phonemes. This is a prerequisite to obtain a reliable estimate for the channel. At the same time, it also assures the nonsingularity of the Hessian matrix in practice. In online applications, channel re-estimation is also postponed until the optimal state sequence over T past frames becomes independent of the current state in the decoder, which in practice means after three to five words. The updating of the channel is illustrated in Fig. 7 for the noisy sentence *"However investment income which represents thirteen percent of the industry's revenues rose eleven percent in the quarter reflecting gains from the rising stock market"* of the Aurora4 database mixed with airport noise at 10-dB SNR. The initial channel was estimated as the log-spectral mean over 200 speech frames. The figure shows the log-Mel value of the 5th, 10th, 15th, and 20th frequency bin for the channel estimate $\boldsymbol{h}$ and their time of update. For example, the initial channel was updated at frame 330, i.e., when the first part of the sentence is recognized. The symbol # indicates that the Viterbi path corresponding to the preceding words is fixed, i.e., it has become independent of the rest of the sentence.

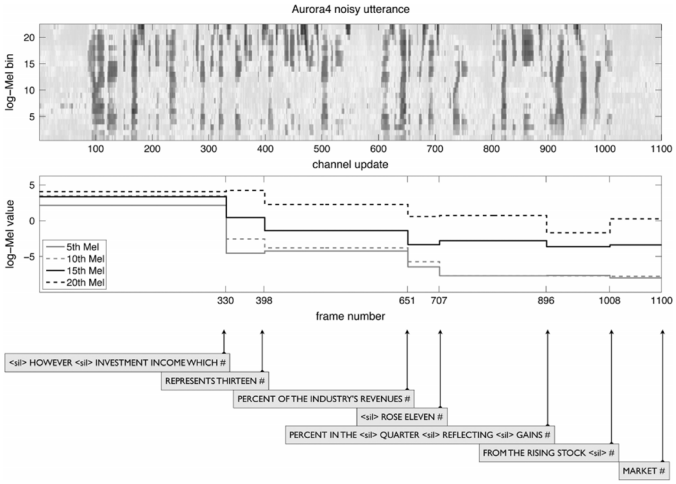

Fig. 7. Illustration of the channel updates.

Since the optimization function $L(\boldsymbol{h})$ exploits the clean speech back-end model, the channel estimation method produces, if we disregard the influence of the noise on $\hat{\boldsymbol{s}}_t$ via masking errors, a negligible bias arising from the background noise. Experimental evidence for the success of this method will be shown in Section VI.

VI. Experiments

In this section, we will first give a qualitative comparison between the different MLE-based imputation techniques that were described in Section III. Therefore experiments will be conducted on the TI-Digits speech database of Aurora-2 for the four noise types of test set A. The MFT-recognizer is based on the HTK software package version 2.2 from Entropic [39]. The digits are modeled as whole word left-to-right HMMs with 16 states per digit and 20 Gaussians with diagonal covariance matrix per state. The optional inter-word silence is modeled by one or three states with 36 Gaussians per state, while leading and trailing silence have three states. The total number of Gaussians is 3628. The ProSpect features were obtained by performing the transformation matrix of (18) to the log-Mel spectral features. Therefore, a Mel-scaled filter bank with $D = 23$ was used identical to the Aurora WI-007 implementation and $K = 3$ order cepstra are used in the ProSpect feature vector.

The baseline results that were obtained by this recognizer without the use of any noise compensation method, are shown in Fig. 8. The graph indicates the recognition accuracy averaged over the four noise types of Aurora2 test set A for different levels of SNR. From this figure it can be seen that it is beneficial to use diagonal GMMs in the cepstral domain (here $K = 13$) rather than in the log-spectral domain. Only at low SNRs, the ProSpect models show to be less accurate than cepstral models. However, this difference in performance does not compensate for the reduction in the computational load that is achieved by ProSpect features in MFT and the performance at low SNR-levels is such that it cannot be considered in the practical employment of the recognizer anyway. Therefore, the use of ProSpects will be preferred over cepstral features in all other experiments in this paper.

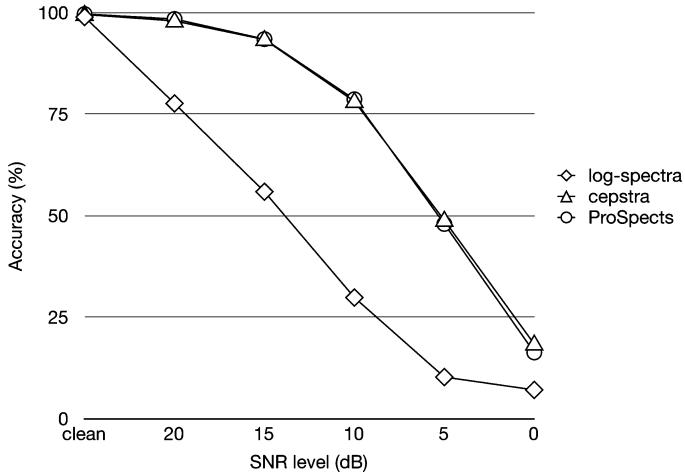

Fig. 8. Baseline results (averaged over the noise types) on the Aurora2 test set A obtained by the recognizer using log-spectral, cepstral and ProSpect models.

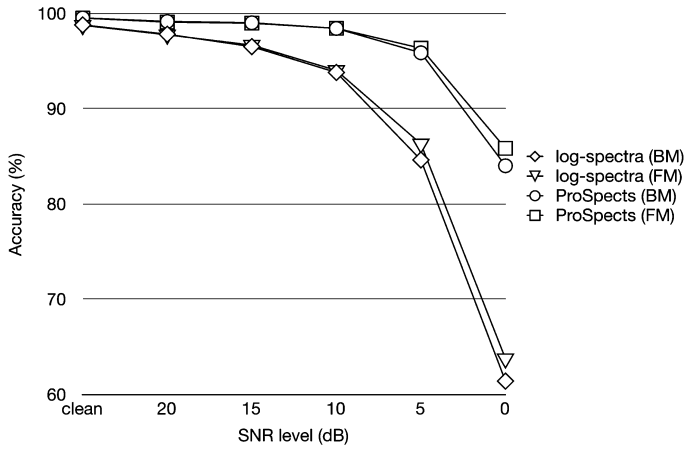

Fig. 9. Performance comparison (averaged over the noise types) on the Aurora2 test set A for the MFT-based recognizer working in the log-spectral and ProSpect domain using binary (BM) or fuzzy (FM) oracle masks.

In Fig. 9, the accuracy of a MFT-based recognizer working in the log-spectral and ProSpect domain is illustrated for the same test set. To ignore the effect of masking errors and to demonstrate the potential of the MFT-based recognizer, *a priori* binary (BM) and fuzzy masks (FM) were used. The dynamic mask for velocity and acceleration features was computed as explained in Section II-A. Besides the significant improvement over the baseline results, the graph also shows that the MLE-based data imputation with ProSpect features always outperforms the imputation in the log-spectral domain.

Experiments were also conducted on the Aurora4 large vocabulary database, derived from the WSJ0 Wall Street Journal 5k-word dictation task [40]. For each of the 2×7 test sets, all 330 utterances from eight different speakers, with a SNR-level that ranges from 5 to 15 dB, are evaluated. Test sets 01–07 were constructed by adding seven types of noise to the Sennheiser close-talking microphone signal: no noise (set 01), car (set 02), babble (set 03), restaurant (set 04), street (set 05), airport (set 06), and train (set 07). Test sets 08-14 were obtained by adding these same noise types to recordings made with 18 different types of microphones. The clean-condition training set contains 7138 utterances from 83 speakers, which is equivalent to 14

hours of speech data. All recordings are made with the close talking microphone and no noise is added.

The MFT-based recognizer was built by adding the required MFT modifications (Section II) to the speaker-independent large vocabulary continuous speech recognition (LVCSR) system that has been developed by the ESAT speech group of the K. U. Leuven. See [41] and [42] for a detailed description of the system. This recognizer was preferred because of its fast experiment turnaround time and good baseline accuracy.

In the MFT-based recognizer as depicted in Fig. 2, the 22-channel Mel filter bank spectra are transformed to the ProSpect domain, where they are modelled with 4961 tied states in the cross-word context-dependent models and with an average of 200 Gaussians with diagonal covariance matrix per state. A bigram language model for a 5k-word closed vocabulary is provided by Lincoln Laboratory, while decoding is done with a time-synchronous beam search algorithm. To compensate for the channel mismatch, the MLE-channel estimation method of Section V has been integrated into the recognizer. The recognition performance will be expressed in terms of the word error rate (WER) which is defined as the number of word errors, i.e., insertions, deletion, and substitution errors, divided by the total number of words in the reference transcription.

Real masks are produced by the MFD which exploits the VQ masking strategy that was introduced in [35] and which can deal with additive and convolutional noise distortions. Here, the key idea was to estimate masks by making only weak assumptions about the noise, while relying on a strong model for the speech. The speech model exploited by the MFD uses the *a priori* knowledge of the human voice, such as harmonicity, voicing, onset, and a vector quantizer to confine the spectral shape of speech signals to a constrained subspace. The involved VQ-codebook was trained on features extracted from the clean speech training set of Aurora4. The number of codebook entries was limited to 500. During speech frames, the decoding seeks to recover the original speech vector from the stored codewords. This way, an estimate for the speech and the noise is generated, which are then substituted in the masking decision criterion (6) or (7) to produce a binary or fuzzy mask vector. To compensate for linear channel distortions, the VQ-based MFD self-adjusts its codebook to the channel during online recognition. More details about the VQ-masks can be found in [35].

Without any noise and channel compensation method, the baseline recognizer runs in real time on a Dual Core AMD Opteron Processor (2.4 GHz/1 MB L2 Cache). In case of the optimization strategy (with 2 iteration steps) using binary VQ-masks for the statics and their derivatives for velocity and acceleration features, all applied in the ProSpect domain, the MFT-based recognizer runs 22 times slower than real time. If fuzzy masks are applied to compensate for the static features, the recognizer is a factor 26 slower than real time. With simple pruning techniques, the computational time of the MFT-recognizer was reduced to a factor 12 and 14, respectively. Therefore, the iterative process of Sections IV-A1 and IV-A2 was gradually stopped for the Gaussians with too small a posterior probability. From experiments not reported in this paper, the pruning parameters were tuned such that computational load is maximally reduced without considerably affecting the recognition performance (an absolute WER increase of 1% was maximally allowed).

Table II presents the word error rate for the MFT-based recognizer where the binary and fuzzy VQ-masks are used for the static features. The masking parameters $\theta = 6$ dB and $\rho = 1.0$ are fixed for all test sets. Masking errors due to incorrectly classifying noise-distorted regions as speech have a higher detrimental effect on the performance of the recognizer than wrongly masking reliable speech regions. This is a first effect that explains why a higher masking threshold (than oracle masks) is desirable for most real masking strategies. A second explanation arises from the fact that the VQ-based MFD underestimates the noise with respect to (w.r.t.) the speech. To compensate for the dynamic features, ternary masks were derived from the binary masks as was described in Section II-A. These results are compared with the baseline, e.g., when no explicit noise and channel reduction algorithm are applied. As can be seen from the table, the results of the VQ-based masks are significantly better than the baseline results. Further potential of the MFT-approach is shown by applying oracle masks ($\theta = -3$ dB) where the *a priori* knowledge of speech and noise are exploited in the decision criterion (6) or (7). From the results in the table, we cannot conclude that an optimization strategy using fuzzy masks performs better than the one using binary masks. The reason is that the fuzzy mask has been derived from the binary masks instead of taking more probabilistically founded decisions about the reliability of the data. However, it illustrates that fuzzy masking strategies can be dealt by the proposed MFT-based speech recognizer without a loss in accuracy due to a different optimization function. Finally, the performance was compared with the results obtained by using model-based feature enhancement (MBFE) of [13] as a noise reduction technique in the front-end. Here, the *a priori* knowledge of the speech and the background noise is incorporated by exploiting a HMM with Gaussian observation probabilities for the clean speech features and for the noise features. The MBFE-algorithm uses a speech model with 256 fully connected Gaussians trained on the clean speech, a 1-Gaussian noise model trained on each noise type individually and an adaptive convolutional noise removal approach. Another difference is that the acoustic models are expressed in the cepstral domain rather than the ProSpect domain. From these results, we conclude that the MFT-based recognizer using VQ-masks has a comparable performance as MBFE without requiring important assumptions about the noise type [35].

Experimental evidence for the effectiveness of the MLE-channel compensation method is given in Fig. 10. Here, the word error rates for each of the 14 test sets of the Aurora4 are presented when the channel is estimated by the technique of Section V and using binary oracle masks. As a reference, the recognition results using log-Mel MS as a channel compensation technique are included. Here, the channel is estimated as the mean over the first 200 speech frames. This is in fact also the initial channel estimate for the MLE-channel compensation method. The second reference experiment can be described as a "cheating" log-Mel MS method since the channel is estimated from all the speech frames of the noise-free version of the speech (i.e., before the artificial addition of the noise).

TABLE II
WORD ERROR RATES (IN %) ON THE AURORA4 DATABASE WITHOUT NOISE COMPENSATION, MFT USING *A PRIORI* (AP) AND VQ-MASKS (VQ) FOR THE STATICS AND THEIR DERIVED TERNARY MASK FOR THE DYNAMIC FEATURES. THE WER OF THE OPTIMIZATION STRATEGY USING BINARY (BM) AND FUZZY (FM) STATIC MASKS ARE BOTH SHOWN. THESE RESULTS ARE COMPARED WITH THE ONES OBTAINED BY A MBFE APPROACH TO COMPENSATE FOR NOISE DISTORTIONS

AURORA4, 16kHz, clean condition training.										
			Close Talk							
test set			01	02	03	04	05	06	07	mic1
baseline			6.82	12.98	32.62	40.95	38.50	32.51	38.37	28.96
MFT	AP	BM	6.54	8.82	13.19	14.72	15.69	11.15	15.69	12.26
		FM	6.54	8.74	13.19	14.22	15.79	11.53	15.49	12.21
	VQ	BM	6.91	11.64	20.44	28.73	24.79	20.16	24.34	19.57
		FM	7.02	11.40	19.24	29.63	24.75	20.42	25.24	19.67
MBFE			5.19	8.61	21.13	29.35	23.41	23.20	22.70	19.08
			Far Talk							
test set			08	09	10	11	12	13	14	mic2	Avg.
baseline			21.35	30.51	47.71	53.26	55.52	46.95	55.41	44.39	36.68
MFT	AP	BM	14.48	18.81	28.06	29.96	31.70	24.29	31.40	25.53	18.89
		FM	15.04	19.02	26.04	28.38	30.95	24.30	29.68	24.77	18.49
	VQ	BM	15.39	22.38	35.59	40.78	40.59	33.72	39.38	32.55	26.06
		FM	14.96	21.71	33.08	40.31	39.83	33.01	38.37	31.61	25.64
MBFE			17.41	24.94	38.45	43.38	42.01	40.80	39.83	35.26	27.17

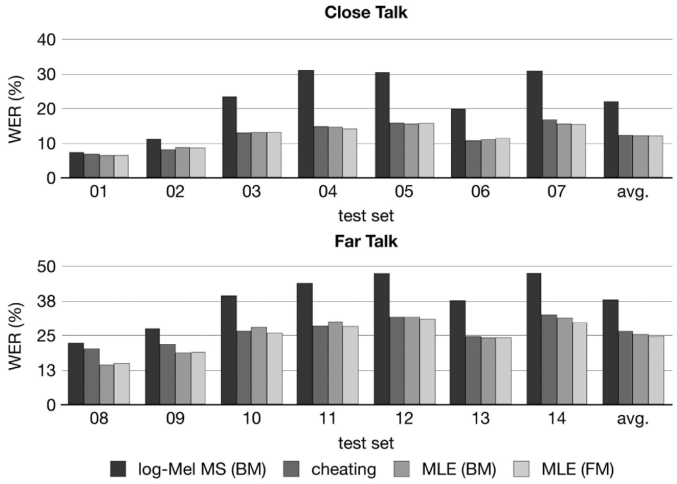

Fig. 10. Word error rate (in %) on the Aurora4 test sets using MFT with binary oracle masks (BM) and using log-Mel MS, cheating log-Mel MS or the MLE-based approach of Section V to compensate for the channel. For the latter approach, the results obtained by the fuzzy oracle mask (FM) are also included.

The comparison of this method with the proposed MLE-based method shows that the latter produces a channel estimation that is hardly biased by the background noise. This can be explained by the fact that the channel is estimated from the back-end model trained on clean speech.

VII. CONCLUSION

Missing feature techniques were applied to large vocabulary speech recognition to jointly compensate for additive and convolutional noise. To this end, MFT first requires a mask estimation procedure in the front-end to estimate the reliability of each spectral component of the noisy input speech. To incorporate the reliability information, the evaluation of the acoustic model in the back-end needs to be modified. Therefore, we described a per-Gaussian-based data imputation technique where an estimate for the clean speech is computed from the information provided by the mask while using the acoustic model in the back-end as a prior and from which the Gaussian likelihood is evaluated. This MFT-technique can be applied in any feature domain that is a linear transform of log-spectra such as cepstra. However, since this requires the optimization of a non-negative least squares problem, ProSpect features were defined to reduce the computational complexity while maintaining the accuracy. Two variants for the imputation technique were formulated to deal with either hard decisions (binary masks) or soft decisions (fuzzy masks) about the reliability of each spectral component. Moreover, this technique was extended to cope with ternary masks to apply MFT on dynamic feature vectors. To compensate for channel distortions, a maximum-likelihood estimation approach was integrated in the MFT-based recognizer to estimate the channel offset by maximizing the log-likelihood of the optimal state sequence of the observation data.

The results on the Aurora4 large vocabulary database significantly improve the baseline result and the MFT-based recognizer has a comparable performance as a MBFE without taking any *a priori* knowledge of the noise into account. Experimental evidence was also given for the effectiveness of the MLE-based approach of Section V as a channel compensation method. By comparing with a cheating log-Mel MS method, we have shown that the estimate of the channel is hardly affected by the background noise.

In future research, we would like to improve the recognition performance of the MFT-based recognizer by increasing the accuracy of the masking decisions to narrow the performance gap with oracle and real masks. Therefore, we will aim to exploit more refined models of the human voice while minimally exploiting the knowledge about the noise. So far, fuzzy masks were constructed as a probabilistic version of the binary masks. It would be interesting to explore masking strategies that estimate more appropriate probabilities for each spectral component of the noisy input speech. Further work is also required to reduce the computational requirements of MFT. Therefore, a faster implementation of the optimization algorithm of Section III is the most important consideration.

APPENDIX I
DERIVATION OF THE GRADIENT AND HESSIAN IN THE MLE-BASED CHANNEL COMPENSATION METHOD

To derive an expression for the gradient and the Hessian of the cost function $L(\boldsymbol{h})$ in (33), we first define $L_t(\boldsymbol{h},\boldsymbol{s}_t)|_{\hat{\boldsymbol{s}}_t}$ as the individual cost contribution for the Gaussian at time t in (33) evaluated in the log-Mel spectral speech estimate $\hat{\boldsymbol{s}}_t$, such that

$$L(\boldsymbol{h})|_{\hat{\boldsymbol{s}}} = \sum_{t=1}^{T} L_t(\boldsymbol{h},\boldsymbol{s}_t)\Big|_{\hat{\boldsymbol{s}}_t}. \tag{35}$$

The gradient of $L_t(\boldsymbol{h},\boldsymbol{s}_t)$ w.r.t. $\boldsymbol{h}$ at $\hat{\boldsymbol{s}}_t$ is then given as

$$\nabla L_t(\boldsymbol{h},\boldsymbol{s}_t)|_{\hat{\boldsymbol{s}}_t} = \frac{\partial L_t(\boldsymbol{h},\boldsymbol{s}_t)}{\partial \boldsymbol{h}}\Big|_{\hat{\boldsymbol{s}}_t} + \left(\frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{h}}\Big|_{\hat{\boldsymbol{s}}_t}\right)' \frac{\partial L_t(\boldsymbol{h},\boldsymbol{s}_t)}{\partial \boldsymbol{s}_t}\Big|_{\hat{\boldsymbol{s}}_t} \tag{36}$$

with

$$\begin{aligned}\frac{\partial L_t(\boldsymbol{h},\boldsymbol{s}_t)}{\partial \boldsymbol{h}}\Big|_{\hat{\boldsymbol{s}}_t} &= -\frac{\partial L_t(\boldsymbol{h},\boldsymbol{s}_t)}{\partial \boldsymbol{s}_t}\Big|_{\hat{\boldsymbol{s}}_t} \\ &= -\boldsymbol{P}_t(\hat{\boldsymbol{s}}_t - \boldsymbol{\mu}_t - \boldsymbol{h}) = -\boldsymbol{g}_t\end{aligned} \tag{37}$$

Here, $\hat{\boldsymbol{s}}_t$ is computed from the constrained minimization of (10) when binary masks are used or from (12) in case of fuzzy masks. While optimizing $\hat{\boldsymbol{s}}_t$, some of the inequality constraints will be active, i.e., the feasible $\hat{\boldsymbol{s}}_t$ that minimizes the cost functions of (10) or (12) lies on that boundary (equality); others will be inactive (strict inequality). Active inequality constraints therefore become equality constraints. Each equality constraint defines a hyperplane (a $D-1$-dimensional space) described by its normal $\boldsymbol{a}_i$. Geometrically, $\hat{\boldsymbol{s}}_t$ is the point on the intersection of all hyperplanes that minimizes $L(\boldsymbol{h})$. Hence, $\boldsymbol{g}_t$ must be perpendicular to all these hyperplanes, for if it would have a nonzero projection in any plane, $\hat{\boldsymbol{s}}_t$ would not minimize $L(\boldsymbol{h})$ subject to the constraints. Therefore, $\boldsymbol{g}_t \in \text{Span}\,(A_t)$ where $A_t = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{M_t}]$($M_t$ is Gaussian dependent) or $A_t^{\perp\prime}\boldsymbol{g}_t = 0$ with $A_t^{\perp}$ a orthonormal matrix perpendicular to A_t. To find an expression for $\partial \boldsymbol{s}_t/\partial \boldsymbol{h}$ evaluated in the point $\hat{\boldsymbol{s}}_t$, assume that $\boldsymbol{h}$ changes with $\Delta\boldsymbol{h}$, then $\hat{\boldsymbol{s}}_t$ changes with $\Delta\hat{\boldsymbol{s}}_t$ and $\boldsymbol{g}_t$ with $\Delta\boldsymbol{g}_t$ such that with (37)

$$\boldsymbol{A}_t^{\perp\prime}\Delta\boldsymbol{g}_t = \boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t(\Delta\hat{\boldsymbol{s}}_t - \Delta\boldsymbol{h}) = 0. \tag{38}$$

We also know that $\hat{\boldsymbol{s}}_t$ has to move in the constraint hyperplane or $\boldsymbol{A}_t'\Delta\hat{\boldsymbol{s}}_t = 0$. Hence, there must exist a vector $\boldsymbol{x}_t$ which satisfies

$$\Delta\hat{\boldsymbol{s}}_t = \boldsymbol{A}_t^{\perp}\boldsymbol{x}_t. \tag{39}$$

After substitution of (39) in (38), we get

$$\boldsymbol{x}_t = (\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\boldsymbol{A}_t^{\perp})^{-1}\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\Delta\boldsymbol{h} \tag{40}$$

and

$$\frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{h}}\Big|_{\hat{\boldsymbol{s}}_t} = \lim_{\Delta\boldsymbol{h}\to 0}\frac{\Delta\hat{\boldsymbol{s}}_t}{\Delta\boldsymbol{h}} = \boldsymbol{A}_t^{\perp}(\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\boldsymbol{A}_t^{\perp})^{-1}\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t. \tag{41}$$

This yields

$$\begin{aligned}\nabla L_t(\boldsymbol{h},\boldsymbol{s}_t)|_{\hat{\boldsymbol{s}}_t} = -(\boldsymbol{P}_t - \boldsymbol{P}_t\boldsymbol{A}_t^{\perp}(\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\boldsymbol{A}_t^{\perp})^{-1}\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t) \\ \times(\hat{\boldsymbol{s}}_t - \boldsymbol{\mu}_t - \boldsymbol{h})\end{aligned} \tag{42}$$

and since $\boldsymbol{A}_t^{\perp\prime}\boldsymbol{g}_t = 0$

$$\nabla L_t(\boldsymbol{h},\boldsymbol{s}_t)|_{\hat{\boldsymbol{s}}_t} = -\boldsymbol{P}_t(\hat{\boldsymbol{s}}_t - \boldsymbol{\mu}_t - \boldsymbol{h}) = -\boldsymbol{g}_t. \tag{43}$$

This result can also intuitively be interpreted as follows: make a perturbation of h_j (the jth component of $\boldsymbol{h}$), then $\hat{\boldsymbol{s}}_t$ must change such that it remains in the hyperplane; hence, $\partial\hat{\boldsymbol{s}}_t/\partial h_j$ lies in all hyperplanes while we know that $\boldsymbol{g}_t$ is perpendicular to these hyperplanes. Hence, the last term in (36) must be zero.

After derivation of $\nabla L_t(\boldsymbol{h},\boldsymbol{s}_t)$ w.r.t. $\boldsymbol{h}$, the Hessian of $L_t(\boldsymbol{h},\boldsymbol{s}_t)$ can be expressed as

$$\begin{aligned}\nabla^2 L_t(\boldsymbol{h},\boldsymbol{s}_t)|_{\hat{\boldsymbol{s}}_t} &= \frac{\partial \nabla L_t(\boldsymbol{h},\boldsymbol{s}_t)}{\partial \boldsymbol{h}}\Big|_{\hat{\boldsymbol{s}}_t} \\ &\quad + \left(\frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{h}}|_{\hat{\boldsymbol{s}}_t}\right)' \frac{\partial \nabla L_t(\boldsymbol{h},\boldsymbol{s}_t)}{\partial \boldsymbol{s}_t}\Big|_{\hat{\boldsymbol{s}}_t} \\ &= \boldsymbol{P}_t - \boldsymbol{P}_t\boldsymbol{A}_t^{\perp}(\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\boldsymbol{A}_t^{\perp})^{-1}\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t.\end{aligned} \tag{44}$$

Note that $\nabla^2 L_t(\boldsymbol{h},\boldsymbol{s}_t)$ is positive semi-definite. Since we know that $\boldsymbol{P}_t$ is symmetric and positive semi-definite (see Section III), we can write $\boldsymbol{P}_t = \boldsymbol{P}_t^{1/2}(\boldsymbol{P}_t^{1/2})'$ and by making use of the QR-decomposition

$$\left(\boldsymbol{P}_t^{1/2}\right)'[\,\boldsymbol{A}_t^{\perp} \quad \boldsymbol{A}_t\,] = \boldsymbol{Q}_t\boldsymbol{R}_t \tag{45}$$

where

$$\boldsymbol{Q}_t\boldsymbol{R}_t = [\,\boldsymbol{Q}_{t,1} \quad \boldsymbol{Q}_{t,2}\,]\begin{bmatrix}\boldsymbol{R}_{t,1} & \boldsymbol{M}\\ 0 & \boldsymbol{R}_{t,2}\end{bmatrix} \tag{46}$$

and with $\boldsymbol{R}_{t,1}$ and $\boldsymbol{R}_{t,2}$ upper triangular matrices. Since the columns of $\boldsymbol{A}_t$ are mutually perpendicular, matrix $[\boldsymbol{A}_t^{\perp} \quad \boldsymbol{A}_t]$ is orthonormal. From (45) and (46) it can be seen that

$$\left(\boldsymbol{P}_t^{1/2}\right)'\boldsymbol{A}_t^{\perp} = \boldsymbol{Q}_{t,1}\boldsymbol{R}_{t,1} \tag{47}$$

and since $\boldsymbol{Q}_t$ is orthonormal, we can write

$$\boldsymbol{Q}_t\boldsymbol{Q}_t' = \boldsymbol{Q}_{t,1}\boldsymbol{Q}_{t,1}' + \boldsymbol{Q}_{t,2}\boldsymbol{Q}_{t,2}' = \boldsymbol{I}_D. \tag{48}$$

After substituting (47) and (48) in (44), the Hessian $\nabla^2 L_t(\boldsymbol{h},\boldsymbol{s}_t)|_{\hat{\boldsymbol{s}}_t}$ can be written as

$$\nabla^2 L_t(\boldsymbol{h},\boldsymbol{s}_t)|_{\hat{\boldsymbol{s}}_t} = \boldsymbol{P}_t^{1/2}\boldsymbol{Q}_{t,2}\boldsymbol{Q}_{t,2}'\left(\boldsymbol{P}_t^{1/2}\right)'. \tag{49}$$

From

$$\begin{aligned}\boldsymbol{Q}_{t,2}'\left(\boldsymbol{P}_t^{1/2}\right)'[\boldsymbol{A}_t^{\perp}\boldsymbol{A}_t] &= [\,\boldsymbol{Q}_{t,2}'\boldsymbol{Q}_{t,1} \quad \boldsymbol{Q}_{t,2}'\boldsymbol{Q}_{t,2}\,] \\ &\quad \times\begin{bmatrix}\boldsymbol{R}_{t,1} & \boldsymbol{M}\\ 0 & \boldsymbol{R}_{t,2}\end{bmatrix} = [\,0 \quad \boldsymbol{R}_{t,2}\,]\end{aligned} \tag{50}$$

we can derive the expression

$$\begin{aligned}\boldsymbol{Q}'_{t,2}\left(\boldsymbol{P}_t^{1/2}\right)' &= [\,0 \quad \boldsymbol{R}_{t,2}\,]\,[\,\boldsymbol{A}_t^{\perp} \quad \boldsymbol{A}_t\,]' \\ &= \boldsymbol{R}_{t,2}\boldsymbol{A}'_t \end{aligned} \tag{51}$$

such that (49) can efficiently be computed as

$$\nabla^2 L_t(\boldsymbol{h}, \boldsymbol{s}_t)|_{\hat{\boldsymbol{s}}_t} = \boldsymbol{A}_t \boldsymbol{R}'_{t,2}\boldsymbol{R}_{t,2}\boldsymbol{A}'_t. \tag{52}$$

After summation of (43) and (52) over all dominant Gaussians of the optimal path, the final expressions of the gradient and the Hessian of $L(\boldsymbol{h})$ for computing the channel estimate in (34) are now given as

$$\begin{aligned}\nabla L(\boldsymbol{h})|_{\hat{\boldsymbol{s}}} &= -\sum_{t=1}^{T} \boldsymbol{P}_t(\hat{\boldsymbol{s}}_t - \boldsymbol{\mu}_t - \boldsymbol{h}) \\ \nabla^2 L(\boldsymbol{h})|_{\hat{\boldsymbol{s}}} &= \sum_{t=1}^{T} \boldsymbol{A}_t \boldsymbol{R}'_{t,2}\boldsymbol{R}_{t,2}\boldsymbol{A}'_t. \end{aligned} \tag{53}$$

Note that the conditions for T that we have formulated in Section V also assure the nonsingularity of the Hessian matrix in practice.

REFERENCES

[1] G. Davis, *Noise Reduction in Speech Applications*, ser. The Electrical Engineering and Applied Signal Processing Series. Boca Raton, FL: CRC, 2002.

[2] L. Josifovski, "Robust automatic speech recognition with missing and unreliable data," Ph.D. dissertation, Univ. of Sheffield, Sheffield, U.K., 2002.

[3] B. Raj and R. Stern, "Robust automatic speech recognition with missing and unreliable acoustic data," *Signal Process. Mag.*, vol. 22, no. 2, pp. 101–116, 2005.

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

[5] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," in *Proc. Eurospeech*, Genua, Italy, Sep. 1991, pp. 1367–1370.

[6] P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view. Application to speech processing in car noise environments," *Speech Commun.*, vol. 12, no. 3, pp. 277–288, Jul. 1993.

[7] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, pp. 117–132, Aug. 1998.

[8] J. Barker and M. Cooke, "Modelling the recognition of spectrally reduced speech," in *Proc. Eurospeech*, 1997, pp. 2127–2130.

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

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

[11] S. V. Vaseghi and B. P. Milner, "Noise-adaptive hidden Markov models based on Wiener filters," in *Proc. Eurospeech*, 1993, pp. 1023–1026.

[12] 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.

[13] V. Stouten, "Robust Automatic Speech Recognition In Time-Varying Environments," Ph.D. dissertation, K. U. Leuven, Leuven, Belgium, Sep. 2006.

[14] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[15] K. Hermus, P. Wambacq, and H. Van Hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP J. Appl. Signal Process. Special Iss. Adv. in Subspace-Based Tech. for Signal Process. Commun.*, vol. 2007, no. 1, pp. 195–204, 2007.

[16] P. Moreno, B. Raj, and R. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. ICASSP*, Atlanta, GA, May 1996, pp. 733–736.

[17] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP*, Albuquerque, NM, Apr. 1990, pp. 845–848.

[18] M. Gales, "Model-based techniques for noise robust speech recognition," Ph.D. dissertation, Univ. of Cambridge, Cambridge, U.K., Sep. 1995.

[19] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[20] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[21] 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.

[22] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Commun.*, vol. 45, no. 1, pp. 5–25, 2005.

[23] L. Josifovski, M. Cooke, P. Green, and A. Vizinho, "State based imputation of missing data for robust speech recognition and speech enhancement," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 2837–2840.

[24] B. Raj, M. L. Seltzer, and R. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 275–296, 2004.

[25] H. Van Hamme, "Robust speech recognition using missing feature theory in the cepstral or LDA domain," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 3089–3092.

[26] H. Van Hamme, "PROSPECT features and their application to missing data techniques for robust speech recognition," in *Proc. Interspeech*, Jeju Island, Korea, 2004, pp. 101–104.

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

[28] A. Morris, J. Barker, and H. Bourlard, "From missing data to maybe useful data: Soft data modelling for noise robust ASR," in *Proc. WISP-01*, Stratford-upon-Avon, U.K., Apr. 2001, pp. 153–164.

[29] F. Faubel, J. McDonough, and D. Klakow, "Bounded conditional mean imputation with Gaussian mixture models: A reconstruction approach to partly occluded features," in *Proc. ICASSP*, Taipei, Taiwan, Sep. 2009, pp. 3869–3872.

[30] M. Van Segbroeck and H. Van Hamme, "Robust speech recognition using missing data techniques in the prospect domain and fuzzy masks," in *Proc. ICASSP*, Las Vegas, NV, Apr. 2008, pp. 4393–4396.

[31] M. Van Segbroeck and H. Van Hamme, "Handling convolutional noise in missing data automatic speech recognition," in *Proc. Interspeech*, Pittsburgh, PA, Sep. 2006, pp. 2526–2565.

[32] P. Renevey and A. Drygajlo, "Missing feature theory and probabilistic estimation of clean speech components for robust speech recognition," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 2627–2630.

[33] M. L. Seltzer, B. Raj, and R. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 379–393, 2004.

[34] H. Van Hamme, "Handling time-derivative features in a missing data framework for robust automatic speech recognition," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 293–296.

[35] M. Van Segbroeck and H. Van Hamme, "Vector-Quantization based mask estimation for missing data automatic speech recognition," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 910–913.

[36] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunneling: Tracking nonstationary noises during speech," in *Proc. Eurospeech*, Aalborg, Denmark, Sep. 1999, pp. 437–410.

[37] C. Cerisara, S. Demange, and J.-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Computer, Speech, Lang.*, vol. 21, no. 3, pp. 443–457, Jul. 2007.

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

[39] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book—Ver. 2.2. 1999, Entropic.

[40] N. Parihar and J. Picone, "Analysis of the aurora large vocabulary evaluations," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 337–340.

[41] "ESAT-PSI Speech," Group [Online]. Available: http://www.esat.kuleuven.be/psi/spraak/

[42] "SPRAAK: Speech processing, recognition and automatic annotation kit," [Online]. Available: http://www.spraak.org/

Maarten Van Segbroeck (S'07–M'09) received the M.Sc. degree in electrical engineering and the Ph.D. degree in engineering from Katholieke Universiteit Leuven, Leuven, Belgium, in 2005 and 2010, respectively.

He joined the research group Processing Speech and Images (PSI), Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, in August 2005. Since January 2010, he has been a Research Assistant in the Institute for the Promotion of Innovation through Science and Technology, Flanders, Belgium (I.W.T.- Vlaanderen). His main interests are robust speech recognition, computational auditory scene analysis, and human–computer interaction.

Hugo Van Hamme (M'92) received the degree of electrical engineer from the Vrije Universiteit Brussels, Brussels, Belgium, in 1987, the M.S. degree from Imperial College, London, U.K., in 1988, and Ph.D. degree from Vrije Universiteit Brussels in 1992.

In 1993, he joined Lernout and Hauspie as a Senior Researcher. Later, he headed the speech recognition research activities in Belgium at this company. In 2001, he joined ScanSoft as a Manager of Research and Engineering for the Automotive Division. Since 2002, he is affiliated full-time as a Professor at Katholieke Universiteit Leuven, Leuven, Belgium. His main research interests are robust speech recognition, computational models of language acquisition, and computer-assisted learning.