Handling Convolutional Noise in Missing Data Automatic Speech Recognition

Maarten Van Segbroeck[*], *Hugo Van hamme*

Katholieke Universiteit Leuven - Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{maarten.vansegbroeck,hugo.vanhamme}@esat.kuleuven.be

Abstract

Missing Data Techniques have already shown their effectiveness in dealing with additive noise in automatic speech recognition systems. For real-life deployments, a compensation for linear filtering distortions is also required. Channel compensation in speech recognition typically involves estimating an additive shift in the log-spectral or cepstral domain. This paper explores a maximum likelihood technique to estimate this model offset while some data are missing. Recognition experiments on the Aurora2 recognition task demonstrate the effectiveness of this technique. In particular, we show that our method is more accurate than previously published methods and can handle narrow-band data.

Index Terms: speech recognition, missing data techniques, convolutional distortion, channel estimation.

1. Introduction

The presence of both additive noise and channel variations results in a decrease in performance of Automatic Speech Recognition (ASR) systems due to the mismatch between the statistics of the resulting input speech and its model, which is often derived under different conditions. In ASR based on Missing Data Techniques (MDT), handling the additive noise involves the estimation of a missing data mask that indicates which spectro-temporal features of the noisy speech are unreliable (missing) or reliable. The latter are treated as clean speech data in the acoustic models of the recognizer's back end. The missing features on the other hand are either marginalized or their value is estimated from the reliable data using the Hidden Markov Model's (HMM) state distribution as a prior (data imputation), see [1]. In the present paper, the data imputation method will be applied.

Channel variations, or convolutional noise, are due to the differences in transmission channels caused by the changes of the distance between mouth and microphone, of the microphone characteristics or of the recording environment. These differences cause a model mismatch between the training and testing conditions. If this mismatch can be described by a linear system with a short impulse response (or smooth transfer function), the mismatch can be modelled by a translation in the cepstral domain. The conventional compensation strategy in ASR is to subtract the cepstral mean from the data, making it insensitive to offsets. Since cepstra and log-spectra are related by a linear transform, this removal of the mean can also be performed on log-spectral features. However, when some log-spectral features are not attributed to speech, but to a different source, as is done in MDT, simple averaging will create an important bias. A method that is compatible with missing data has already been proposed in [2], in which the spectral features are normalized by a factor computed only from the most intense regions of the speech. In the current paper, we present an alternative technique for estimating the channel by a Maximum-Likelihood Estimation (MLE)-based algorithm that updates the initial channel estimate by maximizing the log-likelihood of the optimal state sequence of the observation data.

The outline of this paper is as follows. In section 2 we briefly restate the missing data approach. Section 3 presents a detailed description of the MLE algorithm to compensate for the convolutional noise. An evaluation of the performance of the resulting technique on the Aurora2 connected digit recognition task and the obtained recognition accuracy, can be found in section 4. Conclusions are given in section 5.

2. Maximum likelihood-based imputation in a MDT framework

The speech recognizer is assumed to have a mainstream HMM-based architecture with Gaussian mixture models (GMM). In the front-end, a low resolution MEL-spectral representation is computed by a filter bank with D channels through windowing, framing, FFT and filter bank integration. At frame t, the output of the filter bank with center frequency f will be denoted by $|Y_t(f)|$, $|S_t(f)|$ and $|N_t(f)|$ for the noisy speech, clean speech and noise respectively. The log-MEL-spectral noisy features $\boldsymbol{y}$ are then obtained by stacking $log(|Y_t(f)|)$ for all filter banks in a vector, and likewise for $\boldsymbol{s}$ and $\boldsymbol{n}$. In missing data theory, the following assumption is made for the noisy speech:

$$\boldsymbol{y} \approx max(\boldsymbol{s}, \boldsymbol{n}), \tag{1}$$

where the max-operator works element-wise over the MEL-spectral components. The missing data detector generates a spectral mask that indicates for all t which of the components of $\boldsymbol{y}$ are reliable ($|S_t(f)| \geq |N_t(f)|$) or unreliable ($|S_t(f)| < |N_t(f)|$). In this way, the noisy data vector $\boldsymbol{y}$ is partitioned into a reliable and an unreliable part, $(\boldsymbol{y}_r, \boldsymbol{y}_u)$. The reliable part of $\boldsymbol{s}$ is estimated as $\boldsymbol{y}_r$. In the maximum likelihood per Gaussian-based imputation, the missing part of $\boldsymbol{s}$ is estimated by minimizing the (negative) log-likelihood for each Gaussian mixture component over $\mathbf{s}$:

$$\frac{1}{2}(\boldsymbol{s}-\boldsymbol{\mu})'\boldsymbol{P}(\boldsymbol{s}-\boldsymbol{\mu}) \tag{2}$$

subject to the equality and inequality constraints:

$$\boldsymbol{s}_r = \boldsymbol{y}_r \text{ and } \boldsymbol{s}_u \leq \boldsymbol{y}_u \tag{3}$$

where $\boldsymbol{\mu}$ is the mean of $\boldsymbol{s}$ and $\boldsymbol{P}$ is an inverse covariance or precision matrix. $\boldsymbol{\mu}$ and $\boldsymbol{P}$ are both estimated on clean training data.

[*]This work was supported by "Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen)".

In most MDT systems, GMMs with diagonal covariance in the log-spectral domain are used, resulting in a diagonal structure for $\boldsymbol{P}$ and a tractable MLE for $\boldsymbol{s}$. Higher accuracies are obtained with GMMs with a diagonal covariance in the cepstral domain, in which case $\boldsymbol{P}$ becomes non-diagonal. Imputation then becomes computationally more complex [3]. In this paper, the PROSPECT features defined in [4] will be used, resulting in a known expression for $\boldsymbol{P}$, such that the computational load is reduced while maintaining the accuracy. Despite their performance differences, all these variants of MDT have a known symmetric positive-definite precision matrix $\boldsymbol{P}$. Therefore, the channel compensation method that is proposed in the next section will be generally valid.

3. MDT with channel compensation

3.1. Effect of the convolutional noise

So far, we have only considered the presence of additive noise. In this paper we introduce an extension to the MDT paradigm to remove the convolutional channel distortions as well. In the log-spectral domain, the relationship between the distorted speech vector $\boldsymbol{y}$, the additive noise $\boldsymbol{n}$, the channel $\boldsymbol{h}$ and the clean speech $\boldsymbol{s}$, is given by:

$$\boldsymbol{y} \approx log\big(\exp(\boldsymbol{s}+\boldsymbol{h}) + \exp(\boldsymbol{n})\big) \tag{4}$$

From (4) it is clear that a GMM for $\boldsymbol{s}$ that was trained on undistorted data, can be matched to the distorted data by adding a time-independent shift $\boldsymbol{h}$ to the clean speech means. In the next section, we will derive an expression for this channel shift.

3.2. Channel estimation

The unknown channel parameters are estimated by maximizing the log-likelihood of the optimal state sequence of an observation sequence with length T. This T is chosen dynamically to ensure that we have collected a sufficient amount of speech data. In on-line applications, channel re-estimation is also postponed until the optimal state sequence over T becomes independent of the state in the Viterbi. In this way the optimal state sequence contains a sufficiently large number of Gaussians representative for a diverse set of phonemes, a condition which is fulfilled after the recognition of at least three or four digits in the experiments on the Aurora2 database reported below.

If for a given observation sequence $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T\}$, the optimal state sequence $\hat{Q}$ of the Viterbi decoder is

$$\hat{Q} = arg\max_{Q} P(Q|\boldsymbol{Y}) = \{q_1, q_2, \ldots, q_T\} \tag{5}$$

and if the likelihood of the i-the mixture component of state q with weight $w_{i,q}$ is given by:

$$\xi_{i,q} \sim \frac{\exp\Big(-\frac{1}{2}(\boldsymbol{s}_{i,q}-\boldsymbol{\mu}_{i,q}-\boldsymbol{h})'\boldsymbol{P}_{i,q}(\boldsymbol{s}_{i,q}-\boldsymbol{\mu}_{i,q}-\boldsymbol{h})\Big)}{\sqrt{\det(\boldsymbol{P}_{i,q})}} \tag{6}$$

then the log-likelihood of $\hat{Q}$ is (omitting constant terms):

$$\sum_{t=1}^{T} log\big(\sum_{i=1}^{G} w_{i,q_t}\xi_{i,q_t}\big) \approx \sum_{t=1}^{T} log\big(w_{i_t,q_t}\xi_{i_t,q_t}\big) \tag{7}$$

where i_t is the mixture index of the dominant Gaussian of the mixture of state q_t. Since we now consider only one Gaussian at each time t, we will use the index t in Gaussian variables to indicate the Gaussian (i_t, q_t). The MLE of the channel $\boldsymbol{h}$ can then be obtained by maximizing (7) over $\boldsymbol{h}$ while evaluating each Gaussian t in its optimal point $\hat{\boldsymbol{s}}_t$ according to (2). This is equivalent to the minimization of the cost function L in the points $\hat{\boldsymbol{s}} = \{\hat{\boldsymbol{s}}_1, \hat{\boldsymbol{s}}_2, \ldots, \hat{\boldsymbol{s}}_T\}$:

$$L\mid_{\hat{s}} = \sum_{t=1}^{T} \frac{1}{2}(\hat{\boldsymbol{s}}_t - \boldsymbol{\mu}_t - \boldsymbol{h})'\boldsymbol{P}_t(\hat{\boldsymbol{s}}_t - \boldsymbol{\mu}_t - \boldsymbol{h}) = \sum_{t=1}^{T} L_t\mid_{\hat{s}_t} \tag{8}$$

Note that $\hat{\boldsymbol{s}}_t$ is a function of $\boldsymbol{h}$ and that $L\mid_{\hat{s}}$ depends on the sequence of dominant Gaussians: $\{(i_1,q_1),(i_2,q_2),\ldots,(i_T,q_T)\}$. Hence, iterative optimization is required. Using the Newton-Raphson method, the estimate for the channel $\hat{\boldsymbol{h}}$ can be found as:

$$\begin{aligned}\hat{\boldsymbol{h}} &= \boldsymbol{h} - \Big[\nabla^2 L\mid_{\hat{s}}\Big]^{-1}\cdot\Big[\nabla L\mid_{\hat{s}}\Big] \\ &= \boldsymbol{h} - \Big[\sum_{t=1}^{T}\nabla^2 L_t\mid_{\hat{s}_t}\Big]^{-1}\cdot\Big[\sum_{t=1}^{T}\nabla L_t\mid_{\hat{s}_t}\Big]\end{aligned} \tag{9}$$

This channel update should be applied recursively until convergence, which would imply several recognition passes. Fortunately, experiments (not reported below) have shown that an update strategy with one iteration per T frames suffices. In the next subsections we try to derive an expression for the gradient (∇) and Hessian (∇^2) of L to $\boldsymbol{h}$.

3.2.1. Derivation of $\nabla L\mid_{\hat{s}}$

The gradient of L_t to $\boldsymbol{h}$ is:

$$\nabla L_t\mid_{\hat{s}_t} = \frac{\partial L_t}{\partial \boldsymbol{h}}\mid_{\hat{s}_t} + \Big(\frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{h}}\mid_{\hat{s}_t}\Big)'\frac{\partial L_t}{\partial \boldsymbol{s}_t}\mid_{\hat{s}_t} \tag{10}$$

with

$$\boldsymbol{g}_t = \frac{\partial L_t}{\partial \boldsymbol{h}}\mid_{\hat{s}_t} = \frac{\partial L_t}{\partial \boldsymbol{s}_t}\mid_{\hat{s}_t} = \boldsymbol{P}_t(\hat{\boldsymbol{s}}_t - \boldsymbol{\mu}_t - \boldsymbol{h}) \tag{11}$$

From section 2 we know that $\hat{\boldsymbol{s}}_t$ is chosen such that it minimizes (2) subject to (3). While optimizing $\hat{\boldsymbol{s}}_t$, some of the inequality constraints of (3) will be active, i.e. the feasible $\hat{\boldsymbol{s}}_t$ that minimizes (2) lies on that boundary; others will be inactive. 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. 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 subject to the constraints. Therefore, $\boldsymbol{g}_t \in$ 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 $A_t^{\perp}$ the 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 eq. (11)

$$\boldsymbol{A}_t^{\perp\prime}\Delta\boldsymbol{g}_t = \boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t(\Delta\hat{\boldsymbol{s}}_t - \Delta\boldsymbol{h}) = 0 \tag{12}$$

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

$$\Delta\hat{\boldsymbol{s}}_t = \boldsymbol{A}_t^{\perp}\boldsymbol{x}_t \tag{13}$$

After substitution of (13) in (12), we get

$$\boldsymbol{x}_t = \big(\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\boldsymbol{A}_t^{\perp}\big)^{-1}\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\Delta\boldsymbol{h} \tag{14}$$

and

$$\frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{h}}\mid_{\hat{s}_t} = \lim_{\Delta\boldsymbol{h}\to 0}\frac{\Delta\hat{\boldsymbol{s}}_t}{\Delta\boldsymbol{h}} = \boldsymbol{A}_t^{\perp}\big(\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\boldsymbol{A}_t^{\perp}\big)^{-1}\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t \tag{15}$$

This yields:

$$\nabla L_t = -\Big(\boldsymbol{P}_t - \boldsymbol{P}_t\boldsymbol{A}_t^{\perp}\Big(\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\boldsymbol{A}_t^{\perp}\Big)^{-1}\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\Big)(\boldsymbol{s}_t - \boldsymbol{\mu}_t - \boldsymbol{h}) \tag{16}$$

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

$$\nabla L_t \mid_{\hat{s}_t} = -\boldsymbol{P}_t(\hat{\boldsymbol{s}}_t - \boldsymbol{\mu}_t - \boldsymbol{h}) = -\boldsymbol{g}_t \tag{17}$$

This result can also intuitively be interpreted as follows: make a perturbation of h_i (the i-th component of $\boldsymbol{h}$), then $\hat{\boldsymbol{s}}_t$ must change such that it remains in the hyperplane, hence $\partial\hat{\boldsymbol{s}}_t/\partial h_i$ lies in all hyperplanes while we know that $\boldsymbol{g}_t$ is perpendicular to these hyperplanes. Hence, the last term in eq.(10) must be zero. Finally the expression for $\nabla L \mid_{\hat{s}_t}$ is given by:

$$\nabla L \mid_{\hat{s}} = -\sum_{t=1}^{T}\boldsymbol{g}_t \tag{18}$$

3.2.2. Derivation of $\nabla^2 L \mid_{\hat{s}}$

The Hessian of L_t to $\boldsymbol{h}$ is:

$$\nabla^2 L_t \mid_{\hat{s}_t} = \frac{\partial \nabla L_t}{\partial \boldsymbol{h}} \mid_{\hat{s}_t} + \Big(\frac{\partial \boldsymbol{s}_t}{\partial \boldsymbol{h}} \mid_{\hat{s}_t}\Big)' \frac{\partial \nabla L_t}{\partial \boldsymbol{s}_t} \mid_{\hat{s}_t}$$

$$= \boldsymbol{P}_t - \boldsymbol{P}_t\boldsymbol{A}_t^{\perp}\big(\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t\boldsymbol{A}_t^{\perp}\big)^{-1}\boldsymbol{A}_t^{\perp\prime}\boldsymbol{P}_t \tag{19}$$

Remark that $\nabla^2 L_t$ is positive semi-definite and $\|\nabla^2 L_t\| \leq \|\boldsymbol{P}_t\|$. Since we know that $\boldsymbol{P}_t$ is symmetric and positive semi-definite, we can write $\boldsymbol{P}_t = \boldsymbol{P}_t^{1/2}\big(\boldsymbol{P}_t^{1/2}\big)'$ and by making use of the QR-decomposition:

$$\big(\boldsymbol{P}_t^{1/2}\big)' \cdot \big[\boldsymbol{A}_t^{\perp}\boldsymbol{A}_t\big] = \boldsymbol{Q}_t\boldsymbol{R}_t$$

$$= \big[\boldsymbol{Q}_{t,1} \;\; \boldsymbol{Q}_{t,2}\big] \cdot \begin{bmatrix} \boldsymbol{R}_{t,1} & \boldsymbol{M} \\ 0 & \boldsymbol{R}_{t,2} \end{bmatrix} \tag{20}$$

then eq.(19) can be written as (proof omitted):

$$\nabla^2 L_t \mid_{\hat{s}_t} = \boldsymbol{P}_t^{1/2}\boldsymbol{Q}_{t,2}\boldsymbol{Q}_{t,2}'\big(\boldsymbol{P}_t^{1/2}\big)'$$

$$= \boldsymbol{A}_t\boldsymbol{R}_{t,2}'\boldsymbol{R}_{t,2}\boldsymbol{A}_t' \tag{21}$$

Hence, the expression for the Hessian is given by:

$$\nabla^2 L \mid_{\hat{s}} = \sum_{t=1}^{T}\boldsymbol{A}_t\boldsymbol{R}_{t,2}'\boldsymbol{R}_{t,2}\boldsymbol{A}_t' \tag{22}$$

The conditions for T that we have formulated at the beginning of this section also assure the non-singularity of the Hessian matrix in practice.

4. Experiments

4.1. Corpora, MD recognizer and mask estimation

To evaluate the proposed MLE-based channel compensation technique, we use the Aurora2 TI-Digits speech database, test set A. Since test set A has the same channel characteristics that are used in the training conditions, this test set is regarded as non-distorted (channel 0). Therefore, we have created five additional test conditions. Firstly, the clean speech samples are convolved with the impulse responses of highly distorted channels. The frequency response of these channels are shown in figure 1. Note that channels 4 and 5 are low-pass filters in order to investigate how the system will perform on band-limited data. Subsequently, the four noise types of test set A are added to the filtered clean speech after scaling to the requested SNR (for channels 4 and 5 we have neglected this scaling). The sampling rate is 8kHz. Notice that test set C of the Aurora2 database shows less severe channel mismatch, e.g. it does not contain narrowband speech.

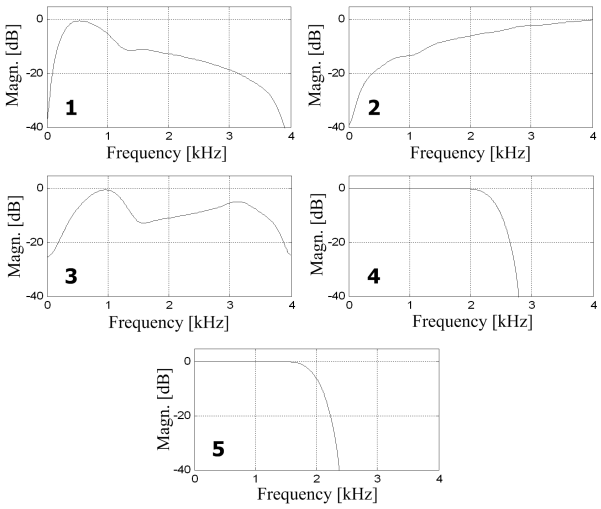

Figure 1: *Frequency responses of the channels used in our experiments.*

The details of the MDT recognizer are described in [5]. In a nutshell, the 23-channel MEL filter bank spectra are transformed to the PROSPECT domain, where they are modelled with 16 HMM states per digit and 20 Gaussians with diagonal covariance per state.

The static missing data are imputed based on two types of masks: *oracle masks* that are derived by comparing the log-spectra of the filtered clean speech and additive noise; *real masks* that are estimated from noisy test set data using harmonicity and SNR information [5]. As motivated in [5], the oracle (real) masks for the dynamic features are derived by applying a delta operator to the oracle (real) masks for the static features.

4.2. Spectral normalization

The MLE-based channel compensation technique is tested against a spectral normalization method based on the ideas of [2]. The spectral features of each MEL-frequency band are normalized by the mean of the N largest features marked as reliable in that frequency band. The normalization factor for the j-the MEL-frequency band is computed as

$$\eta(j) = \frac{1}{N}\sum_{i\in\Gamma(j)} y_r(i,j) \tag{23}$$

where $\Gamma(j)$ is a set containing the indices of the L largest values of the reliable spectro-temporal cells $y_r(i,j)$. The value for N is generally chosen as $N = I/E$ with I the number of frames for an utterance and E is an experimentally derived parameter (here $E = 5$). If this value for N is less than the number of reliable cells, N is set to the number of reliable cells exactly. If no reliable features are available in a frequency band j, no normalization is done, hence $\eta(j)$ is set to 1. Variants were η is computed from the N largest spectro-temporal cells or where the reliable cells are extended with unreliable cells to achieve N data cells for each frequency band, did not result in significant changes in recognition

Oracle masks												
SNR	No comp.				Spectral Normalisation				MLE-based compensation			
(dB)	ch0	ch1	ch2	ch3	ch0	ch1	ch2	ch3	ch0	ch1	ch2	ch3
20 dB	98.86	95.25	93.57	92.69	97.57	95.60	97.42	97.19	99.10	98.51	99.01	99.13
15 dB	98.74	91.64	95.27	90.47	97.33	93.59	97.35	96.49	99.10	96.89	99.03	98.89
10 dB	98.13	84.08	96.10	86.26	96.69	89.09	97.22	94.59	98.52	93.28	98.95	98.04
5 dB	95.61	68.85	94.39	77.19	94.38	79.87	96.11	89.74	96.25	84.41	97.89	94.46
Avg.	97.83	84.96	94.83	86.65	96.49	89.54	97.02	94.50	98.24	93.27	98.72	97.63
Real masks												
SNR	No comp.				Spectral Normalisation				MLE-based compensation			
(dB)	ch0	ch1	ch2	ch3	ch0	ch1	ch2	ch3	ch0	ch1	ch2	ch3
20 dB	98.69	86.78	97.82	86.08	97.67	93.65	94.50	90.93	98.81	96.43	98.85	98.28
15 dB	97.44	78.24	96.46	78.58	96.37	88.86	92.45	88.80	97.90	91.88	97.91	95.82
10 dB	93.95	64.32	92.29	66.72	92.72	79.46	87.00	81.63	94.77	81.74	94.69	87.97
5 dB	83.44	44.44	80.90	51.01	80.27	59.13	73.96	61.96	84.25	62.20	84.81	69.17
Avg.	93.38	68.44	91.86	70.60	91.76	80.28	86.97	80.83	93.93	83.06	94.06	87.81

Table 1: *Average recognition accuracy over the four noise types of Aurora2 test set A for MDT without channel compensation, MDT with spectral normalisation and MLE-based channel compensation; test cases are the filtering characteristics of channels 0-3.*

accuracy. No performance improvements were made when like in cepstral mean subtraction the normalization factor was computed as a geometrical mean.

4.3. Experimental results

Reference results are obtained by applying MDT without channel compensation to the same test set. The mean accuracy over the four noise types of test set A for the channels 0-3 are shown in table 1. These results indicate that the performance increases significantly when a channel compensation method is integrated in a missing data recognition system. Furthermore, table 1 shows that the results of our MLE-based compensation algorithm is superior to those of the spectral normalization method of [2]. The differences in performance between the two methods are most noticeable in the worst distortion conditions (channel 1 and 3) and at low SNRs. Also note that the accuracy for the low distortion cases (channel 0 and 2) sometimes has been worsened with the spectral normalization method, a problem that also has been reported in [2]. From the results of table 1 it is clear that this is not the case for the MLE-based channel compensation method.

The recognition accuracy for the test sets with channels 4 and 5 are shown in table 2. Remark that MDT without compensation has a relatively good performance for these channels. This indicates that the channel is already partially compensated by the imputation of the missing part. However, further increase of the performance is obtained by the extension of the MDT with the MLE-based compensation method.

5. Conclusions

We have presented a new channel compensation method for MDT-based recognizers, where the MLE of the channel is computed from the optimal state sequence of the observation data. For reasons of computational efficiency, we have chosen to work in the PROSPECT domain, but similar results should be obtained if the features are expressed in the cepstral or (log-)spectral domain. Recognition experiments showed the effectiveness of our channel estimation method and that it outperforms the spectral normalization technique. Future work will include the application of our technique to large vocabulary continuous speech recognition.

Oracle masks						
SNR	No comp.		Spec. Norm		MLE	
(dB)	ch4	ch5	ch4	ch5	ch4	ch5
20 dB	95.03	89.93	94.59	90.70	98.46	97.75
15 dB	94.99	90.84	93.88	90.37	97.98	96.91
10 dB	94.23	89.97	92.48	89.30	96.50	94.91
5 dB	90.06	85.29	89.02	85.98	92.30	89.95
Avg.	93.57	89.01	92.49	89.08	96.31	94.88
Real masks						
SNR	No comp.		Spec. Norm		MLE	
(dB)	ch4	ch5	ch4	ch5	ch4	ch5
20 dB	91.40	84.82	96.33	94.51	97.88	96.63
15 dB	90.28	84.71	94.00	91.57	95.57	93.76
10 dB	86.58	81.65	88.46	85.27	90.91	88.44
5 dB	75.08	69.03	74.08	69.73	78.60	74.19
Avg.	85.83	80.05	88.22	85.27	90.74	88.26

Table 2: *Average recognition accuracy over the four noise types of test set A for the filtering characteristics of channels 4-5.*

6. References

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

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

[3] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition,," in *Speech Comm.*, 2004, vol. 43, pp. 275–296.

[4] H. Van hamme, "Prospect features and their application to missing data techniques for robust speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, October 2004, pp. 101–104.

[5] H. Van hamme, "Handling time-derivative features in a missing data framework for robust automatic speech recognition," in *Proc. ICASSP*, Toulouse, France, May 2006.