



# Additive background noise as a source of non-linear mismatch in the cepstral and log-energy domain

Febe de Wet \*, Johan de Veth, Loe Boves, Bert Cranen

*Department of Language and Speech, University of Nijmegen, P.O. Box 9103, Nijmegen 6500 HD, The Netherlands*

Received 21 February 2003; received in revised form 15 October 2003; accepted 15 December 2003

Available online 24 February 2004

---

## Abstract

The aim of this investigation is to determine to what extent automatic speech recognition may be enhanced if, in addition to the linear compensation accomplished by mean and variance normalisation, a non-linear mismatch reduction technique is applied to the cepstral and energy features, respectively. An additional goal is to determine whether the degree of mismatch between the feature distributions of the training and test data that is associated with acoustic mismatch, differs for the cepstral and energy features. Towards these aims, two non-linear mismatch reduction techniques – time domain noise reduction and histogram normalisation – were evaluated on the Aurora2 digit recognition task as well as on a continuous speech recognition task with noisy test conditions similar to those in the Aurora2 experiments. The experimental results show that recognition performance is enhanced by the application of both non-linear mismatch reduction techniques. The best results are obtained when the two techniques are applied simultaneously. The results also reveal that the mismatch in the energy features is quantitatively and qualitatively much larger than the corresponding mismatch associated with the cepstral coefficients. The most substantial gains in average recognition rate are therefore accomplished by reducing training-test mismatch for the energy features.

© 2004 Elsevier Ltd. All rights reserved.

---

## 1. Introduction

In statistical pattern recognition, training data is collected from some population of interest and used to construct models which describe the statistical properties of the individual classes in the

---

\* Corresponding author. Tel.: +31-24-361-5764; fax: +31-24-3612907.

*E-mail addresses:* [f.de.wet@let.kun.nl](mailto:f.de.wet@let.kun.nl) (F. de Wet), [j.deveth@let.kun.nl](mailto:j.deveth@let.kun.nl) (J. de Veth), [l.boves@let.kun.nl](mailto:l.boves@let.kun.nl) (L. Boves), [b.cranen@let.kun.nl](mailto:b.cranen@let.kun.nl) (B. Cranen).

population. New samples may subsequently be classified in terms of their similarity to the models that were constructed during training. This classification strategy is based on the implicit assumption that the training and test samples come from populations with the same or similar statistical properties. A mismatch between the statistical properties of the training and test data violates this assumption. If the assumption is violated, classification performance will deteriorate.

All state-of-the-art automatic speech recognition (ASR) systems are statistical pattern matching machines. Their performance will therefore deteriorate if there is a mismatch between the statistical properties of the training and test data. In the case of ASR, it is often difficult to justify the assumption that training and test tokens are drawn from populations with the same or similar statistical properties. Characteristics of transmission channels and background noise can vary to such an extent that it is often not reasonable to assume that the resulting “speech” signals originate from the same population (statistically speaking). In real-world applications of ASR, there are additional factors that interfere with the assumption of a population with the same underlying statistical properties, such as speaker characteristics, speech style and pronunciation variation. However, in this paper the emphasis will be on the effects of channel distortion and background noise.

Different strategies have been developed to ensure that the statistical properties of training and test data are as similar as possible. One approach that has been investigated in great depth is to condition raw speech signals such that the effects of processes that may significantly alter the statistical properties of the data are eliminated or diminished as much as possible (e.g. Lockwood and Boudy, 1992; Hirsch and Ehrlicher, 1995; Claes and van Compernelle, 1996; Noé et al., 2001). Another approach that has recently been proposed is to use histogram normalisation (HN) to transform the acoustic features that characterise speech signals such that the statistical distributions of the training and test data become as similar as possible. HN is a well-known compensation technique in the field of image processing (e.g. Rus, 1995), but it has also been applied successfully in research on robust ASR (Dharanipragada and Padmanabhan, 2000; Hilger and Ney, 2001; de la Torre et al., 2002). An important advantage of HN is that it can be used in combination with any feature representation. It also does not require any knowledge about the physical causes of the distortions in the data. The idea of normalising parameter transformations is not new and it has been applied widely, even in approaches that started from attempts to remedy the effects of specific physical processes. Mean and variance normalisation (MVN) is probably the best-known technique in this respect (e.g. Cook et al., 1996; Tibrewala and Hermansky, 1997; Viikki and Laurila, 1998).

Because signal conditioning and parameter normalisation operate at different and independent stages of the recognition process, it is possible to combine them. Noise reduction is therefore often followed by mean normalisation, and sometimes also by variance normalisation. Conventional MVN involves a shift of the origin and a linear warping of each feature dimension. In contrast, the approach proposed in Dharanipragada and Padmanabhan (2000); Hilger and Ney (2001); de la Torre et al. (2002) is based on non-linear feature warping. In this paper, we investigate the effects of time domain noise reduction and HN, individually and in combination. In doing so, the main goal is to understand why gains in recognition accuracy are obtained, rather than to improve recognition accuracy as much as possible.

The ASR systems that were used in this study are based on HMMs derived from mel-frequency cepstral coefficients (MFCCs) and an energy parameter ( $\log E$ ). This configuration is generally regarded as state-of-the-art, in ASR research as well as in commercial ASR applications. It should

therefore be possible to integrate our results in most state-of-the-art ASR systems. The combination of MFCCs and  $\log E$  yields feature vectors that describe acoustic signals in terms of the shape of the spectral envelope and the overall energy in the signal at frame level.<sup>1</sup> These two types of features describe different properties of acoustic signals, and it could therefore be expected that they may be affected differently by mismatched training-test conditions.

The strategy for improving robustness followed in this paper is based on the fact that the performance of a statistical pattern recogniser will be optimal if the global distributions of the training and test samples are as similar as possible. Ideally, the match between the distributions of the training and test samples should be normalised independently for all the classes in the data. However, this would require an iterative approach, because individual test tokens would have to be normalised relative to their corresponding class models. Alternatively, one can take a step back and normalise test tokens relative to the overall distribution of the training material, i.e. without distinguishing between the individual classes in the data. This is the approach taken in this paper. The term *overall distributions* is used to refer to the feature distributions that are obtained when the features corresponding to the different speech sound classes are pooled.

Experience has shown that the overall distributions of cepstral parameters derived from clean speech data are approximately normal. In the presence of additive background noise and convolutional channel distortion, the mean values of the overall distributions shift (Junqua and Haton, 1996). In addition, it has been reported that additive noise causes the variances to decrease. It has also been observed that, as SNR decreases, the distributions of the low-order cepstra (especially  $c_1$ ,  $c_2$ , and  $c_3$ ) tend to become bimodal (Openshaw and Mason, 1994). Cepstral mean normalisation (CMN) (e.g. Atal, 1974; Furui, 1981; Mokbel et al., 1994; Rosenberg et al., 1994) and MVN are often used to compensate for the shift in the mean and the reduction in the variance of cepstral parameters derived from noisy speech data.<sup>2</sup> In most practical systems, the mean and variance have to be calculated from a limited amount of data, e.g. only a few data frames are available for mean and variance estimation in real-time applications whereas whole utterances may be used in off-line systems. The compensation accomplished by CMN or MVN is therefore often far from perfect.

The impact of different transmission channels on  $\log E$  features and MFCCs is similar: they cause a constant offset in the long-term mean of the feature tracks. Mean normalisation is therefore conventionally applied to  $\log E$  features in order to limit the extent to which the mean value of the training and test  $\log E$  values may differ. However, the impact of additive noise on the *shape* of the overall distributions of  $\log E$  features is very different from what is usually observed for cepstral features. Contrary to cepstral features, the  $\log E$  features derived from clean speech signals have a bi-modal distribution, with a high-energy component corresponding to voiced speech sounds and a low-energy component corresponding to unvoiced speech sounds, silence and low-intensity non-speech sounds. The impact of additive background noise is most clearly visible

---

<sup>1</sup> This information is usually augmented by the first and second-order regression coefficients of both features. However, the current discussion will concern only the so-called static coefficients. The corresponding regression coefficients will only be derived after an attempt has been made to reduce the statistical mismatch for the static features.

<sup>2</sup> Channel noise can also be compensated for by filtering the time trajectories of acoustic features (e.g. Hirsch et al., 1991; Hermansky and Morgan, 1991; Hermansky and Morgan, 1994; de Veth and Boves, 2003). However, in this study only CMN was used as a means to compensate for convolutional channel distortions.

in the low-energy part of the bi-modal distribution, because it causes an increase in the level of the low-energy parts of the signal. As a consequence, the low-energy component of the overall feature distribution moves towards higher energy levels, which are associated with (voiced) speech sounds. Very high levels of background noise may even result in a uni-modal energy distribution. However, even if the distribution remains bi-modal, background noise will decrease the range from the lowest to the highest  $\log E$  value, and move more probability mass towards higher  $\log E$  values.

While MVN may be used to reduce the differences between the mean values of clean and noisy MFCC and  $\log E$  features and to compensate for the reduction in variance associated with noisy MFCCs, it does not compensate for changes in the shape of the MFCC and  $\log E$  distributions that may occur as a result of additive background noise. Some authors have reported substantial improvements in the recognition rate of a connected digit task if HN is applied to the cepstral coefficients (de la Torre et al., 2002; Segura et al., 2002). In these studies, HN was the only normalising transformation applied to the acoustic parameters (in combination with noise reduction in the form of spectral subtraction).

However, results from other studies have shown that the application of HN to cepstral features yields almost no increase in the recognition accuracy of a continuous speech recognition (CSR) task if it is applied in combination with HN in the mel-filterbank domain (i.e. prior to the application of the discrete cosine transform (DCT)). HN in the mel-filterbank domain can be considered as an alternative to more conventional spectral subtraction (Molau et al., 2001). In the present study, we elaborate on these investigations with the aim to determine to what extent recognition performance may be enhanced if, in addition to the linear compensation accomplished by MVN, a non-linear mismatch reduction strategy is applied to the cepstral and  $\log E$  features, respectively. To that end we compare the effect of a time domain noise reduction (TDNR) scheme (Noé et al., 2001) with HN, and the combination of the two methods.

Since the overall distributions of cepstral and  $\log E$  features are affected differently by additive noise, it is reasonable to expect that non-linear transformations, which aim to equalise the mean and variance as well as the shape of the distributions, will have different effects on cepstral and  $\log E$  parameters. In order to determine whether normalisation of the cepstral and energy features have different effects on recognition accuracy (the single most important criterion in ASR), the compensation strategies were first applied only to the cepstral features, then only to the energy features and finally to both feature types simultaneously.

While features and models can never really be studied in isolation, the focus of this study is on the impact of mismatch reduction in the acoustic signal and feature domains on recognition accuracy. We have therefore chosen to use an experimental design where different transformations are applied to the acoustic signal and the corresponding acoustic features, whereas no changes are made to the recogniser itself. This corresponds to the experimental protocol that was used within the ETSI-Aurora2 experimental framework to evaluate the performance of different ASR systems in noisy conditions (Hirsch and Pearce, 2000). Two sets of experiments were conducted. In the first set, the Aurora2 database and the standard Aurora2 training and evaluation scripts were used. However, the Aurora2 recognition task is limited to connected digit strings. In comparison with the search space described by the phone models, lexicon and language model of a typical CSR system, a digit recognition task based on word models – such as Aurora2 – represents a fairly simple recognition task. In order to determine whether the observations that are made for the Aurora2 experiments

generalise to more complex tasks such as CSR, the experiments were repeated using the CSR engine of an automatic train timetable information system (Strik et al., 1997).

TDNR and HN are described in the next section of this paper. Section 3 gives an overview of the Aurora2 and VIOS databases as well as a description of the experimental set-up. The results and discussions in Section 4 are followed by a general discussion in Section 5. Finally, the conclusions drawn from the outcome of the investigation are presented in Section 6.

## 2. Mismatch reduction techniques

### 2.1. Time-domain noise reduction

The first mismatch reduction technique that was used in this study is the time-domain noise reduction scheme described in Noé et al. (2001); Macho et al. (2002); ETSI (2002). As a first processing step, offset compensation is applied at utterance level. A voice activity detection (VAD) module subsequently classifies each frame as speech or non-speech, based on an estimation of its SNR. The SNR estimate corresponds to the difference between the log-energy spectrum of the current frame and the estimated log-energy spectrum of the noise in the signal. If the VAD module classifies a frame as non-speech, it is used to update the estimate of the noise spectrum. The updated noise spectrum is then used to obtain an estimate of the signal without noise by means of spectral subtraction. The resulting estimates of the noisy and “de-noised” spectra are used to calculate the SNR in each frequency band of the signal. These SNR estimates are subsequently used to derive the transfer function of a Wiener filter. This filter is applied to the noisy signal to obtain a first-pass estimate of the “clean” signal. The filter estimation process is repeated using the estimated noise spectrum and the first-pass estimate of the “clean” signal to obtain a more accurate, second-pass estimate of the Wiener filter. Finally, the “clean” signal is obtained by convolving the original noisy signal with the second-pass Wiener filter in the time domain. For speech data sampled at 8 kHz, the filter estimates are derived from 25 ms frames with a 10 ms frame shift.

### 2.2. Histogram normalisation

As was explained in Section 1, the acoustic mismatch between clean training and noisy test conditions essentially manifests itself as a mismatch between the statistical properties of the training and test data. The aim of HN is to transform the test data such that the match between its overall distribution and that of the training data is improved. When HN is applied to the acoustic features used in speech recognition, it is reasonable to assume that the process which causes the mismatch has an independent effect on the different acoustic vector components. Under this assumption, each feature space dimension may be normalised independently.

The first step in performing HN is to compute the distribution of the training ( $p_k(x)$ ) and test ( $p_k(y)$ ) data for each feature dimension  $k$ . A cumulative distribution density is subsequently derived from both  $p_k(x)$  ( $P_k(x) = \int_{-\infty}^x p_k(x') dx'$ ) and  $p_k(y)$  ( $P_k(y) = \int_{-\infty}^y p_k(y') dy'$ ). Finally, a warping function,  $W_k$ , must be derived such that

$$P_k(x) = W_k[P_k(y)]. \quad (1)$$

HN was implemented according to the methods proposed in Dharanipragada and Padmanabhan (2000); Hilger and Ney (2001); de la Torre et al. (2002). We used 128-bin histograms to approximate  $p_k(x)$  and  $p_k(y)$ .  $p_k(x)$  was calculated using all the training data while  $p_k(y)$  was derived per utterance. In addition, a third-order spline function was used to approximate  $W_k$ . In preliminary experiments, we also investigated the possibility to estimate  $W_k$  using piece-wise linear functions. However, for short utterances the spline function estimates of  $W_k$  yielded better results than the piece-wise linear functions. The minimum and maximum values of  $x_k$  observed in  $p_k(x)$  were used to limit the range of the estimation. Values in the test data that were below the minimum or above the maximum were mapped to  $\min(x_k)$  and  $\max(x_k)$ , respectively.

After  $p_k(x)$  was calculated from the training data, the corresponding acoustic features were also warped according to the function in Eq. (1) at utterance level. This step was taken in order to enforce training-test symmetry in terms of feature transformation. Results from similar studies have shown that the highest recognition rates are obtained if the same feature transformations are applied to the training and test data (Molau et al., 2001).

### 3. Speech data and experimental set-up

#### 3.1. Aurora2

##### 3.1.1. Speech data

The speech data that was used in this study is a subset of the Aurora2 database. The Aurora2 database was derived from a subset of the TI-Digits database (Hirsch and Pearce, 2000). In addition to the original, clean TI-Digits data, it also contains noisy data. The noisy data was created by adding different types of noise to the clean data at different SNRs. The standard Aurora2 experiments include two sets of training data, i.e. clean condition and multi-condition training. The multi-condition training material contains clean data as well as noisy data.

This study involves comparisons in terms of different feature types, different data transformations as well as different recognition tasks. It was therefore decided *not* to use the multi-condition training data because it would have complicated the experimental set-up considerably. Moreover, clean condition training provides a much more challenging mismatch reduction problem than multi-condition training.

Three test sets were defined for the Aurora2 task, i.e. sets A, B, and C. Sets A and B each contain 4004 and Set C 2002 utterances. All three test sets are made up of a mixture of clean and noisy data. The clean and noisy signals were downsampled to 8 kHz and subsequently digitally filtered to simulate the effect of two standard communication channels, i.e. the G.712 and the MIRS (ITU, 1996). The frequency response of the two filters are given in Hirsch and Pearce (2000). According to these figures, the G.712's frequency response is flat between 300 and 3400 Hz, whereas the MIRS' response is sloped, slightly attenuating frequencies below 2000 Hz. Both filters were implemented using the ITU STL96 software package. The speech and noise data were passed through one of the filters before the noise was added to the clean material.

Test sets A and B have the same channel properties as the training data (G.712) but differ from each other in the types of noise they contain. Set A was made using suburban train, babble, car and exhibition hall noise while the noisy data in Set B comprise restaurant, street, airport and

station noise. The suburban train noise from test set A and the street noise from test set B were used to create the noisy data in test set C. In addition, the transmission channel properties of the data in test set C were simulated using the MIRS filter instead of the G.712 filter.

### 3.1.2. Hidden Markov modelling

The reference recognition system that was developed for Aurora2 (Hirsch and Pearce, 2000) was used for all the digit experiments. The system is based on hidden Markov word models and implemented in HTK (Young et al., 1997). Each model has 16 states with a mixture of 3 continuous density Gaussians per state. Model topology only allows left-to-right transitions without skipping states. In addition to the 11 digit models (one, two, three, four, five, six, seven, eight, nine, zero, oh), two silence models were also trained: one corresponding to silences at the beginning and the end of the utterances (three states, six Gaussians per state) and one corresponding to silences between words (single state tied to the middle state of the three-state silence model).

## 3.2. VIOS

### 3.2.1. Speech data

The VIOS database was collected with an on-line version of a spoken dialogue system that provides train timetable information in the Netherlands (Strik et al., 1997). The speech data was recorded over the public switched telephone network. Speaker, handset and channel characteristics are unknown. The language of the corpus is Dutch and the speech is spontaneous and unprepared. A total of 33,471 utterances were collected and transcribed. 25,104 utterances were used for training (83,876 words corresponding to 8.9 h of speech excluding leading, utterance-internal and trailing silence). The remaining 8,358 utterances (28,048 words corresponding to 3.0 h speech) were allocated to the test set.

None of the utterances used for training had a high background noise level. Noisy test conditions were simulated by adding train station noise to the original test utterances. The noise data was collected in the hall of a train station in the Netherlands. The results of informal listening tests indicated that the noise is similar to the Aurora2 train station noise. The noise was added to the original test data such that the resulting acoustic signals had SNRs of 0, 10 and 20 dB, respectively.

In Fig. 1, the VIOS training ( $\circ$ ) and test ( $\times$ ) data are compared with the corresponding data from the Aurora2 database (training ( $\square$ ) and test ( $+$ )). The SNRs shown in the figure were calculated using NIST software (NIST, 1999). The values on the  $x$ -axis correspond to the signals' SNR before the application of TDNR. The corresponding values on the  $y$ -axis were calculated after TDNR had been applied. It should be kept in mind that the values in Fig. 1 represent *estimated* SNRs. For this reason, they are not equal to the nominal SNRs that were used to create the noisy data. Furthermore, the figure is primarily intended as a means to compare the Aurora2 and VIOS data in terms of SNR and not to evaluate the success of the TDNR algorithm to improve SNR.

According to the data in Fig. 1, both the Aurora2 and VIOS clean training and test data are well matched in terms of SNR. However, the clean Aurora2 data has a much higher SNR than the clean VIOS data. The figure also shows that the SNRs of the Aurora2 and VIOS noisy test data

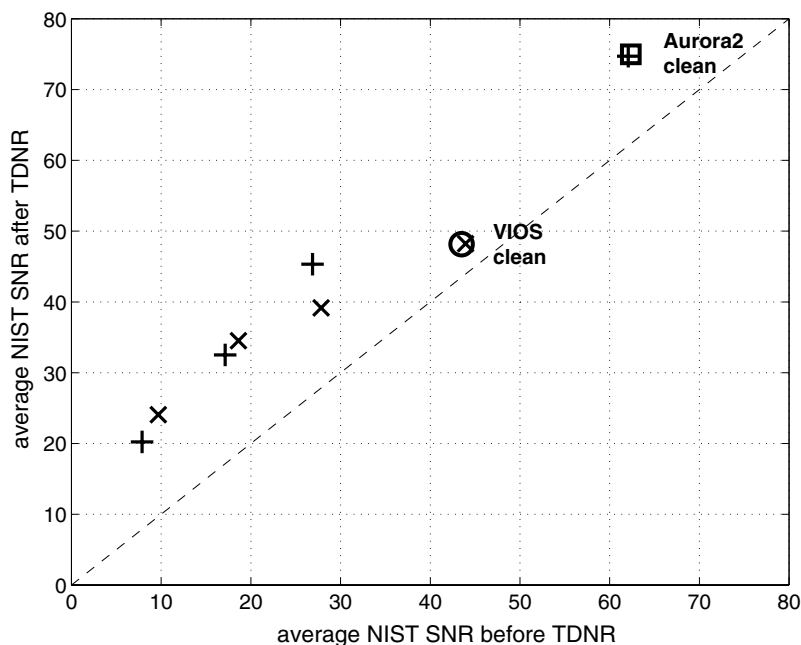


Fig. 1. Mean SNR of the clean and noisy (train station) Aurora2 (training ( $\square$ ), test (+)) and VIOS (training ( $\circ$ ), test ( $\times$ )) data before and after the application of TDNR.

differ only slightly: the SNR of the 0 dB VIOS data is a little higher than the SNR of the corresponding Aurora2 values and TDNR does not improve the SNR of the 20 dB VIOS data to the same extent as for the 20 dB Aurora2 data.

### 3.2.2. Hidden Markov modelling

The Phicos speech recognition software was used to conduct the CSR experiments (Steinbiss et al., 1995). In total, 37 phone models were trained: 33 context-independent and four context-dependent. The four context-dependent models were used to model the two allophones of /l/ and /r/, respectively. The allophones correspond to the pre-vocalic and post-vocalic realisation of the two phones. Two models were also defined to describe noise/non-speech events and silence. The silence model was a single-state HMM. All the other models consisted of six HMM states with states 2, 4 and 6 sharing their emission probability density functions with states 1, 3, and 5, respectively. All the HMMs were left-to-right with only self-loops, transitions to the next state or to the next state plus one. The emission probability density functions were described as a mixture of 32 Gaussian probability density functions (diagonal covariance matrices).

The VIOS training lexicon contained 1,106 words. The acoustic models were initialised using a linear segmentation of the speech portions of the signals, as determined with a silence-speech detector. After initialisation, a fixed number of Viterbi optimisation passes was used to train the models. As a next step, the number of Gaussians per state was doubled. To this aim a K-means clustering algorithm was applied using the segmentations obtained in the previous Viterbi pass (Steinbiss et al., 1995). After splitting, Viterbi optimisation was applied again. For the models

with 1, 2, and 4 Gaussians per state the number of Viterbi optimisation passes was 2, 3, and 3, respectively. For 8 Gaussians and beyond, 7 Viterbi optimisation passes were used.

The recognition lexicon contained 980 words, and 1.2% of the words in the test set were out-of-vocabulary. During recognition the acoustic models were combined with unigram and bigram language models derived from the training data. The average test set perplexity of the recognition task was 36.7.

The test set was divided into a development test set (1/4 of the data) and an independent evaluation test set (the remaining 3/4 of the data). The word entrance penalty (WEP) and the language model factor (LMF) were jointly optimised on the development test set. Recognition performance was subsequently determined using the independent evaluation test set. The WEP and LMF that were found to be optimal for the baseline system were used for all the CSR experiments reported on in Section 4.

### 3.3. Acoustic pre-processing and feature extraction

Fig. 2 gives an overview of the acoustic pre-processing procedure that was used to derive the spectral shape ( $c_1, \dots, c_{12}$ ) and energy ( $\log E$ ) features. The shaded blocks in the figure correspond

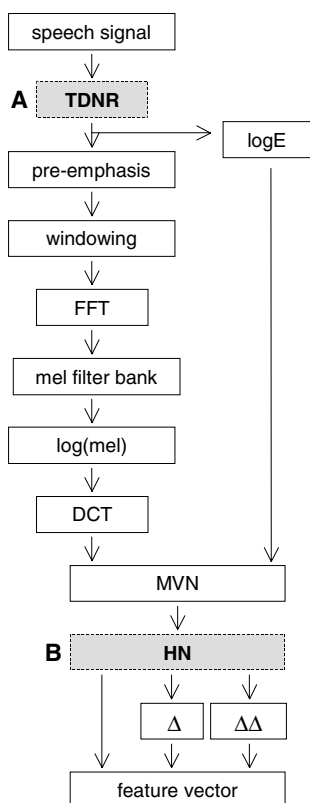


Fig. 2. Schematic overview of the feature extraction and mismatch reduction modules.

to the mismatch reduction techniques described in Section 2. Block A represents TDNR and block B HN.

A pre-emphasis factor of 0.98 and a 25 ms Hamming window shifted with 10ms steps were used to prepare the data for spectral analysis. After a 256-point FFT, 16 mel-scaled log-energy values were calculated for each frame. The filters in the mel bank were triangularly shaped, half overlapping and uniformly distributed on a mel-frequency scale between 122 and 2146 mel, corresponding to 80–4000 Hz on a linear frequency scale. 12 MFCCs were derived from the mel-bank outputs using a DCT. The log of the total energy ( $\log E$ ) was also calculated for each frame.

The MFCC and  $\log E$  values were normalised to have zero mean and unit variance (at utterance level) according to the MVN scheme described in Viikki and Laurila, 1998. After mean and variance normalisation, the first and second-order time derivatives of the resulting features were also computed (using a regression length of 9 in both instances) and included in the acoustic feature vectors. In the experiments where features were transformed using HN, the first and second-order time derivatives were calculated after the application of HN.

### 3.4. Mismatch reduction experiments

Three mismatch reduction experiments were carried out. In Experiment I, only block A in Fig. 2 was included in the acoustic pre-processing, i.e. the training and test data (including the clean signals) were subjected to the time domain noise reduction scheme described in Section 2.1 *before* feature extraction. In Experiment II, only block B was included in the acoustic pre-processing, i.e. HN was applied to the MFCCs and  $\log E$  *after* feature extraction. In Experiment III, both blocks A and B were active, i.e. TDNR was applied *before* and HN *after* feature extraction.

In each of the experiments, the mismatch reduction schemes were implemented in the following order: (1) not at all (baseline); (2) only for the *cepstral features* and using the baseline  $\log E$ ; (3) only for the *energy features* and using the baseline MFCCs; and (4) for *both* the MFCC and the  $\log E$  features. Training/test symmetry was observed in all experiments, i.e. the transformations that were applied to the test data were also applied to the training data.

## 4. Results and discussion

The results in this section are defined in terms of recognition accuracy, i.e.  $(N - S - D - I/N) \times 100\%$ , where  $N$  is the total number of words in the test set,  $S$  denotes the total number of substitution errors,  $D$  the total number of deletion errors and  $I$  the total number of insertion errors. Section 4.1 gives an overview of the results that were obtained for the Aurora2 connected digit recognition experiments. The results for the VIOS CSR experiments are subsequently presented in Section 4.2.

### 4.1. Aurora2

In clean, matched conditions the average recognition accuracy of the Aurora2 recogniser was 99.0% ( $\pm 0.2$ ). None of the mismatch reduction techniques that were applied caused a significant change in the recogniser's ability to classify the clean test material. The recognition accuracies

reported in this section were calculated according to the Aurora2 protocol, i.e. the mean recognition accuracy for each test set was obtained by taking the average of the recognition rates measured in 0, 5, 10, 15, and 20 dB SNR. The values in the columns labelled *Average* were calculated as  $0.4 \times \text{Set A} + 0.4 \times \text{Set B} + 0.2 \times \text{Set C}$ . The weighting factors account for the fact that test set C contains only half as many utterances as test sets A and B. The 95% confidence intervals of the average values are shown in parentheses.

#### 4.1.1. Experiment I: TDNR

The results that were obtained when the Aurora2 data was submitted to TDNR before feature extraction are summarised in Table 1. The values in the table show that, calculating only the MFCCs from the data after TDNR yields a marginal, statistically insignificant increase in the average recognition accuracy. In contrast, the average recognition rate increases substantially when only the  $\log E$  values are derived from the data after TDNR. The best results are obtained when both the MFCCs and  $\log E$  are calculated after the application of TDNR. However, the difference between the last two rows of Table 1 is much smaller than the difference between the last and the second row of the table. This observation suggests that the change in the  $\log E$  feature values accounts for most of the total gain in recognition rate.

#### 4.1.2. Experiment II: HN

Table 2 gives an overview of the results that were obtained for the Aurora2 task after the application of HN in the acoustic feature domain. According to the values in Table 2, the average recognition performance obtained when HN is applied only to the MFCCs does not differ from the baseline. However, applying HN on  $\log E$  yields a marked increase in recognition rate. When both the MFCCs and  $\log E$  are transformed, the largest part of the total gain can therefore be attributed to the transformation applied to  $\log E$ . This trend in the results was also observed in Experiment I. However, the results for test sets A and C differ substantially from those measured

Table 1  
Recognition accuracy for the Aurora2 digit recognition task after the application of TDNR

Transformed features	Set A	Set B	Set C	Average
Baseline	72.1	72.4	74.0	72.6 ( $\pm 0.4$ )
MFCCs	72.5	72.9	73.8	72.9 ( $\pm 0.4$ )
$\log E$	82.3	82.0	79.3	81.6 ( $\pm 0.3$ )
MFCCs and $\log E$	83.3	82.5	79.3	82.2 ( $\pm 0.3$ )

Table 2  
Recognition accuracy for the Aurora2 digit recognition task after the application of HN

Transformed features	Set A	Set B	Set C	Average
Baseline	72.1	72.4	74.0	72.6 ( $\pm 0.4$ )
MFCCs	72.0	72.5	74.1	72.6 ( $\pm 0.4$ )
$\log E$	80.1	81.8	81.7	81.1 ( $\pm 0.3$ )
MFCCs and $\log E$	80.8	82.7	82.3	81.8 ( $\pm 0.3$ )

in Experiment I: the mean recognition accuracy for test set A is almost 3% lower and the mean recognition accuracy for test set C is 3% higher than in Experiment I.

#### 4.1.3. Experiment III: TDNR and HN

The recognition accuracies that were measured when both TDNR and HN were applied to the Aurora2 data are summarised in Table 3. These results show that it is still possible to achieve a substantial improvement in recognition performance if HN is applied in combination with TDNR. Once again, most of the gain can be attributed to the transformation applied to the  $\log E$  feature. However, in contrast to the results of Experiments I and II, transforming only the MFCCs also leads to a statistically significant increase in recognition accuracy.

#### 4.1.4. Discussion

The values in Table 3 compare favourably with those reported by other authors for the Aurora2 task (e.g. Segura et al., 2002; Macho et al., 2002; Adami et al., 2002). However, it should be kept in mind that many of the techniques proposed in these studies include a strategy to remove excessive non-speech frames before recognition takes place, e.g. frame dropping (Macho et al., 2002) or feature vector selection (de Veth et al., 2001b). The results in Table 3 could probably be improved if frame dropping or feature vector selection were applied.

As was pointed out in Section 1, the aim of this investigation is to determine to what extent recognition performance may be enhanced if, in addition to the linear compensation accomplished by MVN, a non-linear mismatch reduction technique is applied to the cepstral and  $\log E$  features, respectively. In addition, we also wanted to determine whether the degree of mismatch between the feature distributions of the training and test data that is associated with acoustic mismatch, differs for the cepstral and energy features.

In Openshaw and Mason (1994), it was reported that, if Gaussian white noise was added to clean acoustic signals, the global distributions of  $c_1$ ,  $c_2$ , and  $c_3$  tended to become bi-modal (non-Gaussian) with decreasing SNR. Fig. 3 shows the overall distributions of  $c_1$ ,  $c_2$ ,  $c_3$  and  $c_4$  derived from the clean training data (solid line) and 0 dB SNR train station test data (dotted line) in the baseline condition (i.e. when only MVN is applied). These distributions are clearly not bi-modal and – with the possible exception of  $c_1$  – there are no substantial differences between the shapes of the distributions corresponding to the noisy test data and those derived from the clean training data. The modal value of noisy  $c_1$  seems to differ from its clean counterpart. Apparently, mean normalisation at utterance level is not able to remove the effects of noise on the central tendency completely. Similar observations were made for other noise types.

One possible explanation for the difference between the shapes of the distributions shown in Fig. 3 and those that were reported on in Openshaw and Mason (1994) is the difference between

Table 3

Recognition accuracy for the Aurora2 digit recognition task after the application of both TDNR and HN

Transformed features	Set A	Set B	Set C	Average
(TDNR) baseline	83.3	82.5	79.3	82.2 ( $\pm 0.3$ )
MFCCs	83.6	82.9	80.1	82.6 ( $\pm 0.3$ )
$\log E$	84.0	83.8	82.7	83.7 ( $\pm 0.3$ )
MFCCs and $\log E$	84.5	84.3	83.3	84.2 ( $\pm 0.3$ )

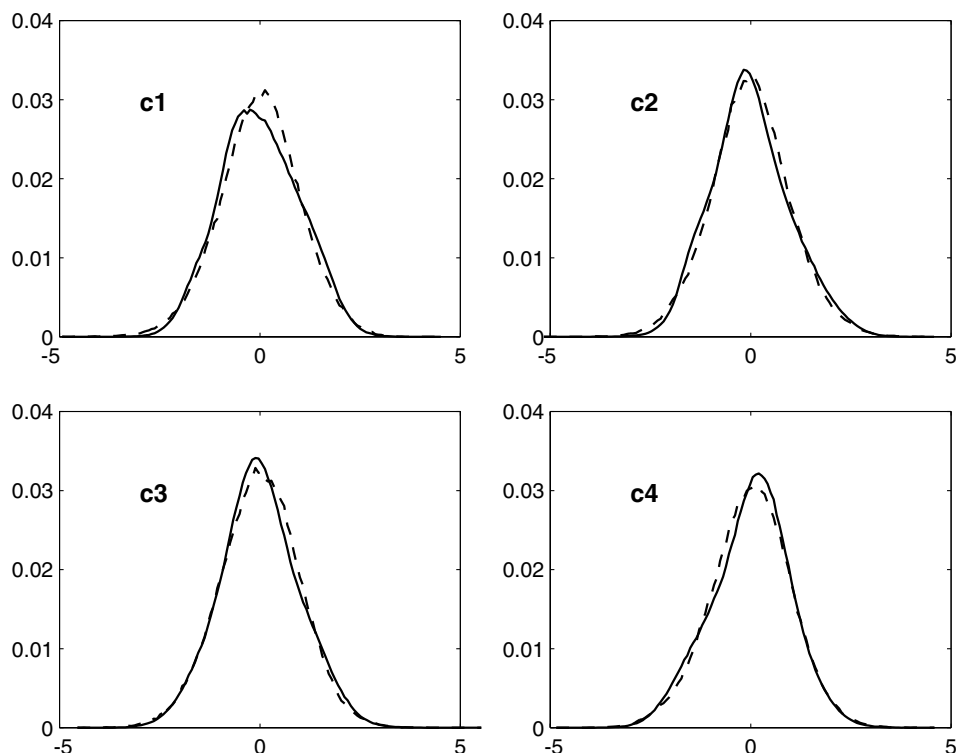


Fig. 3. Overall distribution of  $c_1$ ,  $c_2$ ,  $c_3$  and  $c_4$  derived from clean training data (solid line) and 0 dB SNR train station test data (dotted line).

the effect of Gaussian white noise and the types of noise in the Aurora2 database. Another possible explanation for the observed differences is the fact that in Openshaw's study there were only 20 different speakers in the speaker population whereas speech from 104 different speakers is included in the Aurora2 database. These discrepancies indicate that experimental results may be strongly influenced by the details of the experimental set-up and that observations often do not generalise from one study to the other.

The results that were presented in Tables 1 and 2 revealed that neither the application of TDNR nor HN resulted in a significant increase in the average recognition rate if only the MFCCs were transformed. These results suggest that, in the baseline condition, the application of HN and TDNR does not do much to reduce the mismatch between the overall distributions of the training and test MFCCs. However, the data in Fig. 3 shows that there is little room for improvement, because the overall distributions of the cepstra derived from clean and noisy data are highly similar. In contrast, the recognition rates in Table 3 improved significantly when only the MFCCs were transformed. This observation seems to indicate that, after the application of TDNR, there is still a residual mismatch between the distributions of the training and test data which may be compensated for by the application of HN. Evidence in support of this assumption is provided in Fig. 4.

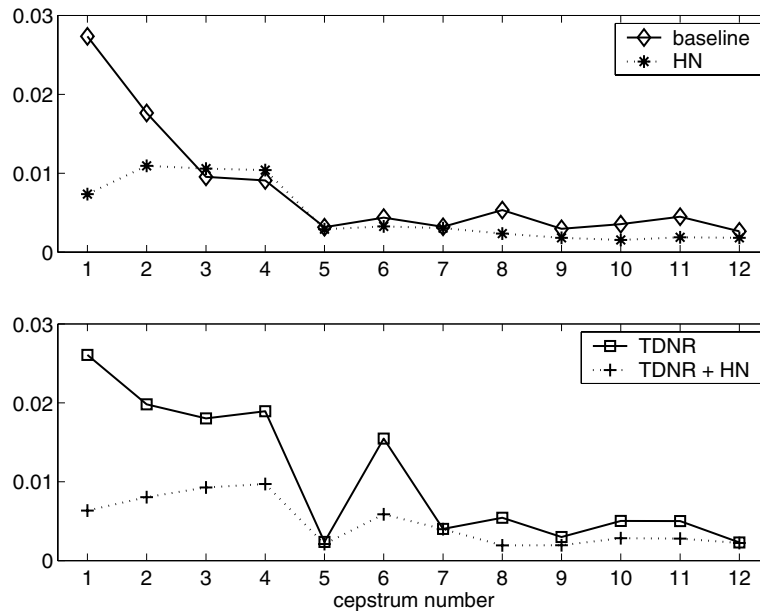


Fig. 4. Kullback divergence between the distributions of the training and test data for the 12 cepstral coefficients. Distances for the baseline condition and after the application of HN. (Top) Distances after TDNR and TDNR + HN have been applied (Bottom).

Fig. 4 shows the distance – in terms of Kullback divergence (Kullback, 1959; Basseville, 1989) – between the global distributions of the clean training and noisy (0 dB, train station) test data for the 12 cepstral coefficients. The trend in Fig. 4 is in good agreement with results from a previous study where it was also observed that the degree of training-test mismatch (measured in terms of Kullback divergence) is much higher for the lower than the higher order cepstra (de Wet et al., 2002). The graph in the top part of the figure shows the Kullback distances corresponding to the baseline condition (◇) and after the application of HN (\*). The data shows that the application of HN reduces the distance between the training and test distributions for  $c_1$  and  $c_2$ , but has very little impact on the distances corresponding to the higher order cepstra. This probably explains why there is almost no difference between the results of the baseline and the HN experiments if only the MFCCs are transformed. The second graph in Fig. 4 shows that, in comparison with the baseline condition (◇), the application of TDNR (□) increases the distances between the overall training and test distributions for  $c_2$ ,  $c_3$ ,  $c_4$  and  $c_6$ . However, a combination of HN and TDNR (+) compensates for this increase to a large extent. The data therefore confirms the idea that the application of TDNR gives rise to statistical mismatch in the cepstral domain, which can be compensated for by HN. Similar observations were made for other noise types. Although the absolute distances are smaller at higher SNRs, the same trend is visible in the data.

In terms of  $\log E$ , the results show that recognition accuracy is significantly enhanced if the energy features are calculated from the data after TDNR, if HN is applied to the  $\log E$  features and if the two techniques are applied simultaneously. The corresponding changes in the overall feature distributions are illustrated in Fig. 5. The topmost distribution in the figure corresponds to the overall distributions of the clean (solid line) and noisy (0 dB, train station)  $\log E$  features. The

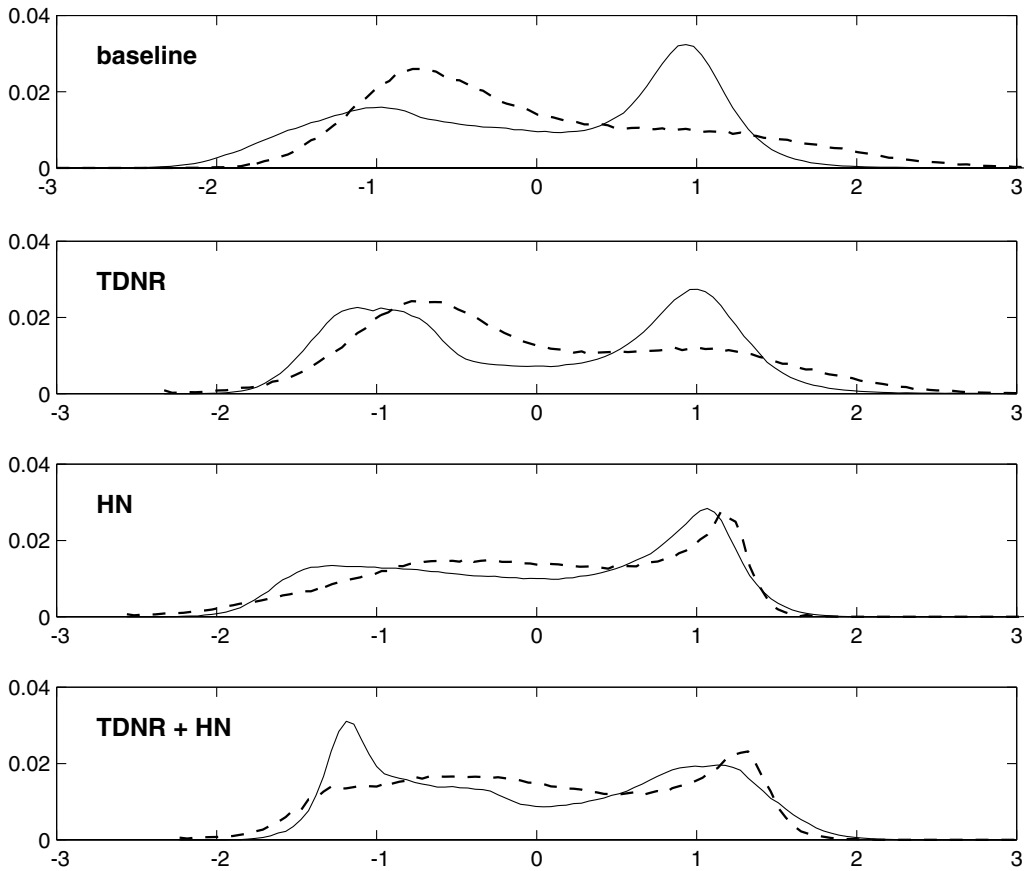


Fig. 5. Overall distribution of  $\log E$  derived from clean training data (solid line) and 0 dB SNR train station test data (dotted line) in the baseline condition, after the application of TDNR, after HN, and after TDNR + HN have been applied.

data in the figure shows that there is a vast mismatch between the shapes of the clean and noisy features' distributions. It also shows that the application of TDNR (second set of distributions), HN (third set), as well as a combination of TDNR and HN (last set), alleviates this mismatch to a large extent. However, despite the reduction in mismatch, the Kullback divergence values corresponding to the distributions in Fig. 5 are still more than an order of magnitude larger than the biggest values in Fig. 4.

In Section 4.1.2, it was observed that there was a large difference between the  $\log E$  transformed and MFCCs and  $\log E$  transformed results obtained with TDNR and HN. The mean recognition accuracy for test set A was almost 3% lower with HN than with TDNR, whereas the mean recognition accuracy for test set C was 3% higher with HN than with TDNR. An analysis of the individual test sets revealed that this difference could be ascribed to the fact that the results for TDNR in the babble, car, and exhibition hall noise in test set A is much better than the corresponding HN results. On the other hand, the HN results for test set C (especially in suburban train noise) are superior to their TDNR counterparts.

These observations suggest that there is an interaction between the noise type, the channel properties of the training and test data, and the mismatch reduction techniques that were used in these experiments. The TDNR algorithm clearly finds it more difficult to compensate for the suburban train noise after the acoustic signals have passed through the channel simulated in test set C than the channel simulated in test set A. According to the noise and channel properties given in Hirsch and Pearce, 2000, the frequency response of the MIRS filter that was used to create the data in test set C is sloped in its pass-band. As a result, the lower frequencies are slightly attenuated with respect to the higher frequencies. In addition, the long-term average spectrum of the suburban train noise has two spectral peaks: one around 400 Hz and one around 2.5 kHz. Passing the suburban train noise through the MIRS filter could therefore enhance the high-frequency components of the noise to such an extent that the efficiency of the TDNR algorithm deteriorates.

In all the experiments that were described in Section 4.1, the gain in average recognition rate accomplished by transforming the  $\log E$  features accounted for most of the overall gain in average system performance. The fact that reduced mismatch in the global distribution of the  $\log E$  features leads to such large improvements in recognition performance may be an artefact of the Aurora2 experimental set-up. Because of the low complexity of the task, the  $\log E$  features may have a larger impact on recognition performance than the MFCCs. In order to determine whether the observations made for the Aurora2 experiments generalise to a more complex task such as CSR, Experiments I, II and III were repeated using the CSR engine of an automatic train timetable information system (Strik et al., 1997). The results that were obtained for these experiments are reported on in the next section.

## 4.2. VIOS

The recognition accuracy of the baseline VIOS system in clean, matched conditions is 88.8% ( $\pm 0.5$ ). This value is in good agreement with results reported by others for the same recognition task (Kessens et al., 1999). The corresponding value for the Aurora2 data is 99.0%. This difference can be accounted for by various factors. In terms of the complexity of the underlying search space, CSR is a much more difficult task than connected digit recognition. Fig. 1 in Section 3.2.1 also showed that the SNR of the clean VIOS data is much lower than the SNR of the clean Aurora2 data. It should also be kept in mind that the effect of the transmission channel was simulated for the Aurora2 data while the VIOS data was recorded over a public switched telephone network. Moreover, within each test set, the same channel properties were simulated for all the Aurora2 data whereas the 33,471 utterances in the VIOS database were recorded over many different transmission channels. As was illustrated in Fig. 2, MVN was performed during acoustic feature extraction. Mean normalisation was applied in order to compensate for the effect of different transmission channels. However, as was pointed out in Section 1, the compensation accomplished by MVN is not always perfect.

In this section, the CSR results that were measured in the noisy test conditions are reported in terms of the average of the recognition accuracies measured in 0, 10, and 20 dB SNR for the VIOS data. By means of comparison, the tables also include the corresponding average values calculated from the Aurora2 recognition rates measured in 0, 10 and 20 dB SNR train station noise. The 95% confidence intervals of the average accuracies are shown in parentheses.

#### 4.2.1. Experiment I: TDNR

After the application of TDNR, the CSR system achieved a recognition rate of 88.6% on the clean data. This value is not significantly different from the baseline (88.8%). Table 4 gives an overview of the results that were obtained for the VIOS data when TDNR was applied to the acoustic signals before feature extraction.

According to the values in Table 4, the VIOS results do not follow exactly the same pattern as the Aurora2 results. In both instances, the best overall results are obtained when both the MFCCs and the  $\log E$  features are calculated after the application of TDNR. However, the corresponding increase in average recognition rate is much smaller for the VIOS CSR task than for the Aurora2 digit recognition task. In addition, when only the MFCCs are calculated from the data after TDNR, it leads to a significant improvement in the average recognition accuracy for the VIOS task. In fact, there is no significant difference between the average results for the *MFCCs transformed* and the *log E transformed* experimental conditions. This is clearly different from the corresponding Aurora2 results.

#### 4.2.2. Experiment II: HN

For the clean VIOS data the recognition rate of the VIOS CSR system dropped from 88.8% to 87.9% after the clean data was transformed using HN. This deterioration in recognition rate is statistically significant. The VIOS results that were obtained in noisy test conditions when HN was applied in the acoustic feature domain are summarised in Table 5.

The trend in the Aurora2 and VIOS results in Table 5 is almost the same, the only difference being that the absolute gain in average recognition rate is much bigger for Aurora2 than for VIOS. If only the MFCCs are transformed, there is no significant change in the average recognition rate. Applying HN only to the  $\log E$  features results in a substantial increase in the average recognition accuracy. If both types of features are transformed, the average system performance improves slightly for both Aurora2 and VIOS. However, the difference between the *log E transformed* and the *MFCCs and log E transformed* results is not significant.

Table 4  
Recognition accuracy after the application of TDNR

Transformed features	Aurora2	VIOS
Baseline	69.8 ( $\pm 1.3$ )	59.4 ( $\pm 0.4$ )
MFCCs	70.2 ( $\pm 1.3$ )	60.6 ( $\pm 0.4$ )
$\log E$	78.4 ( $\pm 1.3$ )	60.8 ( $\pm 0.4$ )
MFCCs and $\log E$	79.3 ( $\pm 1.3$ )	61.6 ( $\pm 0.4$ )

Table 5  
Recognition accuracy after the application of HN

Transformed features	Aurora2	VIOS
Baseline	69.8 ( $\pm 1.3$ )	59.4 ( $\pm 0.4$ )
MFCCs	69.8 ( $\pm 1.3$ )	59.1 ( $\pm 0.4$ )
$\log E$	78.2 ( $\pm 1.3$ )	63.6 ( $\pm 0.4$ )
MFCCs and $\log E$	78.9 ( $\pm 1.3$ )	63.8 ( $\pm 0.4$ )

#### 4.2.3. Experiment III: TDNR and HN

As was observed when HN was applied in isolation, the recognition rate of the VIOS CSR system deteriorated significantly when the clean data was submitted to both TDNR and HN: the recognition accuracy dropped from 88.8% to 87.5%. The results that were measured in the noisy test conditions are summarised in Table 6.

In contrast to the corresponding Aurora2 results, the VIOS data show a significant drop in average recognition rate if only the MFCCs are transformed. However, both systems' average performance is significantly enhanced if the  $\log E$  features are transformed. For Aurora2, the best overall results are obtained when both the MFCCs and  $\log E$  are transformed, but the resulting average recognition rate is not significantly better than when only  $\log E$  is transformed. For the VIOS task, the highest average recognition rate is measured when only  $\log E$  is transformed. The VIOS results deteriorate slightly but not significantly when both the MFCCs and  $\log E$  are transformed.

#### 4.2.4. Discussion

The results that were presented in this section show that not all the observations that were made for the Aurora2 connected digit recognition task generalise unconditionally to the VIOS CSR task. For instance, in the Aurora2 experiments the TDNR results were superior to the HN results (compare Tables 1 and 2). For the VIOS experiments, on the other hand, the application of HN leads to much higher recognition accuracies than the application of TDNR (compare Tables 4 and 5).

Although the application of HN in the cepstral domain did not improve the average Aurora2 results, it also did not hurt system performance. In contrast, applying HN in the cepstral domain in combination with MVN and TDNR caused the average VIOS recognition rates to drop. In the latter case (TDNR + HN), the deterioration in system performance even turned out to be significant. One possible explanation of this effect might be that HN has trouble with short utterances (almost 10% of the VIOS test set consists of the answers to yes/no questions), where very little data is available to estimate the parameters of the warping function. A poorly estimated warping function may result in a distorting rather than a compensating data transformation. However, there was no significant difference between the recognition rates of the single word utterances in the baseline condition and those in the HN and TDNR + HN experiments.

Another reason for the negative impact of HN on recognition performance could be the speech/non-speech ratio in the VIOS database: many of the utterances have long leading and trailing silences. Consequently, more than 50% of the data corresponds to non-speech, which means that non-speech could have a strong influence on the shape of the overall feature distributions. The

Table 6  
Recognition accuracy after the application of both TDNR and HN

Transformed features	Aurora2	VIOS
(TDNR) baseline	79.3 ( $\pm 1.3$ )	61.6 ( $\pm 0.4$ )
MFCCs	79.6 ( $\pm 1.3$ )	60.5 ( $\pm 0.4$ )
$\log E$	81.1 ( $\pm 1.3$ )	64.5 ( $\pm 0.4$ )
MFCCs and $\log E$	81.6 ( $\pm 1.3$ )	64.2 ( $\pm 0.4$ )

results of Experiments II and III suggest that it does more harm than good to derive warping functions for the cepstral features from these distributions.

A striking similarity between the Aurora2 and VIOS results is the significance of the  $\log E$  features: even though the absolute gain in average recognition accuracy is much smaller for the VIOS data, it is clear that reducing the mismatch for the  $\log E$  features significantly improves recognition performance. This observation seems to suggest that well-matched  $\log E$  features are of primary importance for successful speech recognition, even in the more complex search space of a CSR system.

## 5. General discussion

In Segura et al. (2002), the authors reported a large gain in recognition rate for the Aurora2 task if spectral subtraction was applied in the mel-filterbank domain and HN was applied to the resulting cepstral coefficients. No other normalising transforms were applied in the cepstral domain, i.e. HN was used to accomplish mean, variance, and “shape” normalisation. In contrast, the results reported in Molau et al. (2001) showed almost no improvement in the recognition performance of a CSR task when spectral subtraction in the mel-filterbank domain (implemented by means of HN) was applied in combination with HN in the cepstral domain. However, in Molau et al. (2001) MVN was applied to the cepstral coefficients before performing HN. The difference between the outcomes of these two studies can probably be explained by the fact that in Segura et al. (2002) normalisation in the cepstral domain was lumped into a single transformation step, whereas in Molau et al. (2001) normalisation in the cepstral domain was performed sequentially. The results of the current study are in good agreement with those reported in Molau et al. (2001): after cepstral mean subtraction and variance normalisation, a non-linear transformation of cepstral coefficients does not improve recognition performance.

The data in Fig. 3 provide a possible explanation for these observations. After mean and variance normalisation there is not that much left to compensate for in the cepstral domain: the overall distributions of the cepstra derived from the clean training data are uni-modal and fairly symmetric (approximately normal) and this shape does not change substantially in the presence of additive background noise. This observation seems to confirm results from a previous study where it was found that various types of additive background noise result in small, almost equal distortions in all MFCCs (de Veth et al., 2001a). Even noise types that are (approximately) band-limited, e.g. car noise, are “smeared out” over the entire cepstral feature vector by the application of the DCT. In that study, it was found that a robust distance function which was able to enhance system performance in the mel-filterbank domain was not able to compensate for the effect of the smaller, less obvious distortions in the corresponding cepstral features. The two mismatch reduction techniques that were used in the current investigation do not seem to be able to do so either. Even though TDNR and HN are both fairly successful in dealing with the quantitatively and qualitatively large mismatch observed for the  $\log E$  features, their application in the cepstral domain has almost no influence on recognition performance.

The results in Section 4 show that, if HN is applied *only* to cepstra after mean subtraction and variance normalisation, there is almost no change in the recognition accuracy of the baseline Aurora2 system. Moreover, the same feature transformation resulted in deteriorated results

for the CSR experiments. A substantial increase in recognition rate was only observed when one of the non-linear mismatch reduction techniques was applied to  $\log E$ . These observations suggest that the gain reported in Segura et al. (2002) may probably be attributed to linear channel compensation in the cepstral domain and that the non-linear distortions remaining after spectral noise reduction are probably limited to the distributions of the  $\log E$  features. It is also possible that the application of spectral subtraction introduced statistical mismatch in the cepstral feature distributions similar to the mismatch that was observed for TDNR in this study (cf. Fig. 4). In that case, the application of HN in Segura et al. (2002) may have helped to alleviate the impact of the residual mismatch in the cepstral domain as it did in Experiment III (cf. Section 4.1.3). However, mismatch in the  $\log E$  features is likely to have had just as strong an influence on the results reported in Segura et al. (2002) as on the results reported in the current study.

Almost all the mismatch reduction techniques that were investigated in this study yielded substantial improvements in *average* recognition accuracy. However, in most cases the *average* recognition rate is “boosted” by large gains in very low SNR conditions. The fact that the increase in recognition accuracy is moderate – and often not even significant – in higher SNR conditions is obscured by comparing the systems in terms of *average* recognition rate. This observation is illustrated by the data in Fig. 6 where recognition accuracy is plotted as a function of SNR.

A significant increase in recognition accuracy (relative to the baseline condition) is indicated with a “+” in the corresponding bar, a significant deterioration with a “–” and an empty bar indicates no significant difference. The results in Fig. 6 show that, for both Aurora2 and VIOS, the biggest absolute gain in recognition accuracy is at 0 dB. While all three mismatch reduction techniques are able to enhance recognition rates significantly at 0 and 10 dB, only two of them are successful in the 20 dB condition for the Aurora2 task and they all fail to increase recognition performance significantly in the 20 dB VIOS experiments. Fig. 6 also shows that, for the Aurora2 experiments, the technique that yields the best recognition performance in the 0 dB test condition (TDNR + HN) does not lead to a significant improvement in system performance in the 20 dB test condition. The VIOS results show a similar trend, with the additional disadvantage that the combination of TDNR and HN leads to a significant drop in recognition accuracy for the clean data. Despite the substantial gain in recognition rate that is achieved in the 0 dB condition, one could argue about the usefulness of this result: in some applications it may be useful to understand a little more than nothing at all. In those instances, the techniques that were investigated in this study may produce useful results. However, they will not be of much use in situations where “clean condition” recognition accuracies are required. On the other hand, if a system is designed to be used almost only in “clean” conditions, it is probably not worth the effort to make it as robust as possible to additive background noise – especially not if the transformations that are applied to enhance noise robustness leads to deteriorated recognition performance in the clean condition. Application-specific considerations should therefore be taken into account in the design of robust ASR systems.

The results that were obtained for HN in this study may be improved by deriving class-specific warping functions instead of global warping functions. Classes may be words or phones, depending on the system architecture. Further research is required to determine how many frames are required for a reliable estimate of the test data distributions. A buffer of adequate length can then be created to collect frames and the warping function can subsequently be updated whenever the buffer is full. This strategy differs from the adaptation methods that are currently used in ASR

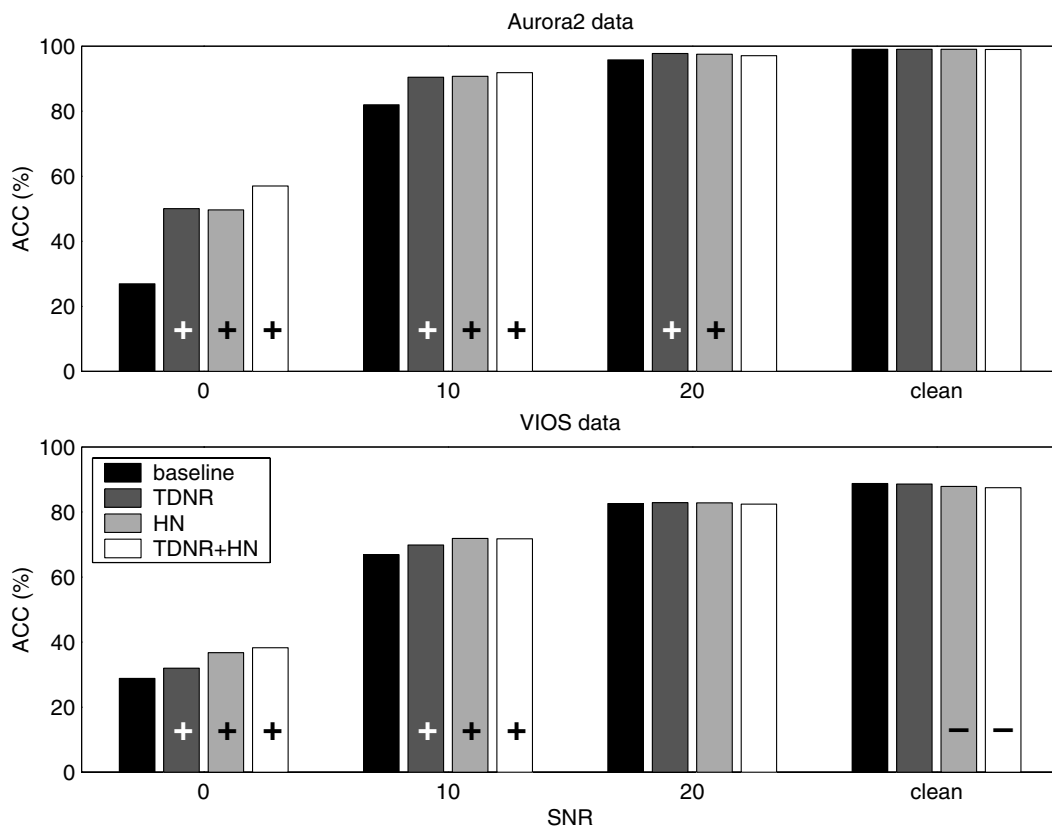


Fig. 6. Recognition accuracy as a function of SNR for the Aurora2 digit recognition task (top) and the VIOS CSR task (bottom). (significant increase (+), significant deterioration (-), no significant difference (empty bar) – relative to the baseline (filled bar)).

in that it is a *normalisation* transformation in the acoustic *feature* domain and not an *adaptation* of the acoustic *models*. However, the proposed non-linear, class-dependent normalisation would require a two-pass recognition: one pass to determine a tentative class-specific segmentation and another after the application of HN. In off-line applications, e.g. automatic transcription tasks, the corresponding gain in recognition rate may be worth the effort.

## 6. Conclusions

The results that were presented in this study show that recognition performance may be enhanced if, in addition to the linear compensation accomplished by MVN, non-linear mismatch reduction techniques such as TDNR and HN are applied. The results also reveal that there is a huge difference in the degree of mismatch between the feature distributions of the training and test data for cepstral and  $\log E$  features. The mismatch between the  $\log E$  distributions is quantitatively and qualitatively much larger than the corresponding mismatch associated with the MFCC

distributions. The biggest gains in average recognition rate are therefore accomplished by reducing the training-test mismatch in the  $\log E$  feature distributions.

The results of the Aurora2 experiments indicate that, after the application of MVN, the global distributions of cepstral coefficients are inherently robust to the impact of additive background noise. In the presence of additive background noise, there were no substantial changes in the shape of the overall feature distributions, not even at low SNRs. On the other hand, the distributions of the  $\log E$  features were shown to be extremely vulnerable to additive background noise. The mismatch reduction techniques that were used in this study could compensate for the resulting training-test mismatch to a certain extent.

Not all the observations that were made for the Aurora2 data generalise to the VIOS CSR task. The absolute gain in recognition rate that could be accomplished by the mismatch reduction techniques is smaller for VIOS than for Aurora2. The results also do not exhibit exactly the same trends, e.g. TDNR outperformed HN in the Aurora2 experiments whereas HN achieved much better results for the VIOS task (on average). For both the Aurora2 and the VIOS experiments the best recognition rates were obtained when TDNR and HN were used in tandem. For these experiments (i.e. after the application of TDNR), the best Aurora2 results were achieved when HN was applied to both the MFCC and the  $\log E$  features, whereas the best VIOS results were obtained when HN was applied only to the  $\log E$  features. In both the Aurora2 and VIOS experiments, the major part of the total gain in system performance could be attributed to a reduction in the mismatch between the overall distributions of the  $\log E$  coefficients, despite the difference in the complexity of the underlying search spaces.

## References

- Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., Sivasdas, S., 2002. QUALCOMM-ICSI-OGI features for ASR. In: *Proceedings of ICSLP 2002*, pp. 21–24, Denver, CO, USA.
- Atal, B., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America* 55, 1304–1312.
- Basseville, M., 1989. Distance measures for signal processing and pattern recognition. *Signal Processing* 18, 349–369.
- Claes, T., van Compernelle, D., 1996. Spectral estimation and normalisation for robust speech recognition. In: *Proceedings of ICSLP 1996*, pp. 1997–2000, Philadelphia, PA, USA.
- Cook, G.D., Christie, J.D., Clarkson, P.R., Hochberg, M.M., Logan, B.T., 1996. Real-time recognition of broadcast radio speech. In: *Proceedings of ICASSP 1996*, vol. I, pp. 141–144, Atlanta, GA, USA.
- de la Torre, A., Segura, J.C., Benitez, M.C., Peinado, A.M., Rubio, A.J., 2002. Non-linear transformation of the feature space for robust speech recognition. In: *Proceedings of ICASSP 2002*, Orlando, FL, USA.
- de Veth, J., Boves, L., 2003. On the efficiency of classical RASTA filtering for continuous speech recognition: keeping the balance between acoustic pre-processing and acoustic modelling. *Speech Communication* 39, 269–286.
- de Veth, J., de Wet, F., Cranen, B., Boves, L., 2001a. Acoustic features and a distance measure that reduce the impact of training-test mismatch in ASR. *Speech Communication* 34 (1–2), 57–74.
- de Veth, J., Mauuary, L., Noé, B., de Wet, F., Siemel, J., Boves, L., Jouviet, D., 2001b. Feature vector selection to improve ASR robustness in noisy conditions. In: *Proceedings of Eurospeech 2001*, pp. 201–204, Aalborg, Denmark.
- de Wet, F., de Veth, J., Cranen, B., Boves, L., 2002. Accumulated Kullback divergence for the analysis of ASR performance in the presence of noise. In: *Proceedings of ICSLP 2002*, pp. 1069–1073, Denver, CO, USA.
- Dharanipragada, S., Padmanabhan, M., 2000. A nonlinear unsupervised adaptation technique for speech recognition. In: *Proceedings of ICSLP 2000*, vol. IV, pp. 556–559, Beijing, China.

- ETSI, 2002. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. Available from <<http://pda.etsi.org/pda/queryform.asp>> (search on 'aurora').
- Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-29*, 254–272.
- Hermansky, H., Morgan, N., 1991. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In: *Proceedings of Eurospeech 1991*, pp. 1367–1370, Genova, Italy.
- Hermansky, H., Morgan, N., 1994. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing* 2, 578–589.
- Hilger, F., Ney, H., 2001. Quantile based histogram equalisation. In: *Proceedings of Eurospeech 2001*, pp. 1135–1138, Aalborg, Denmark.
- Hirsch, H.-G., Ehrlicher, C., 1995. Noise estimation techniques for robust speech recognition. In: *Proceedings of ICASSP 1995*, pp. 153–146, Michigan, USA.
- Hirsch, H.-G., Meyer, P., Ruehl, H., 1991. Improved speech recognition using high-pass filtering of subband envelopes. In: *Proceedings of Eurospeech 1991*, pp. 413–416, Genova, Italy.
- Hirsch, H.-G., Pearce, D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: *Proceedings of ASR 2000*, pp. 181–188, Paris, France.
- ITU, 1996. ITU recommendation G.712: "Transmission performance characteristics of pulse code modulation channels". Available from <<http://www.itu.int/>>.
- Junqua, J.-C., Haton, J.-P., 1996. On the use of a robust speech representation. In: *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic Publishers, Boston, USA.
- Kessens, J.M., Wester, M., Strik, H., 1999. Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Special issue of Speech Communication: Modeling Pronunciation Variation for Automatic Speech Recognition* 29 (2–4), 193–207.
- Kullback, S., 1959. *Information Theory and Statistics*. Wiley, New York.
- Lockwood, P., Boudy, J., 1992. Experiments with a nonlinear spectral subtractor (NNS) hidden Markov models and the projection, for robust speech recognition in cars. *Speech Communication* 11, 215–228.
- Macho, D., Mauuary, L., Noé, B., Cheng, Y.-M., Ealey, D., Jouvét, D., Kelleher, H., Pearce, D., Saadoun, F., 2002. Evaluation of noise-robust DSR front-end on Aurora databases. In: *Proceedings of ICSLP 2002*, pp. 17–20, Denver, CO, USA.
- Mokbel, C., Pachès-Leal, P., Jouvét, D., Monné, J., 1994. Compensation of telephone line effects for robust speech recognition. In: *Proceedings of ICSLP 1994*, pp. 987–990, Yokohama, Japan.
- Molau, S., Pitz, M., Ney, H., 2001. Histogram normalisation in the acoustic feature space. In: *Proceedings of ASRU 2001*, Madonna di Campiglio, Trento, Italy.
- NIST, 1999. NIST SNR evaluation software. Available from <<http://www.itl.nist.gov/iaui/894.01/>>.
- Noé, B., Siemel, J., Jouvét, D., Mauuary, L., Boves, L., de Veth, J., de Wet, F., 2001. Noise reduction for noise robust feature extraction for distributed speech recognition. In: *Proceedings of Eurospeech 2001*, pp. 433–436, Aalborg, Denmark.
- Openshaw, J.P., Mason, J.S., 1994. Optimal noise-masking of cepstral features for robust speaker identification. In: *ESCA Workshop on automatic speaker recognition, identification and verification*, pp. 231–234, Martigny, Switzerland.
- Rosenberg, A., Lee, C.-H., Soong, F., 1994. Cepstral channel normalization techniques for HMM-based speaker verification. In: *Proceedings of ICSLP 1994*, pp. 1835–1838, Yokohama, Japan.
- Rus, J.C., 1995. *The Image Processing Handbook*. CRC Press, Boca Raton, FL.
- Segura, J.C., Benitez, C., de la Torre, A., Rubio, A., 2002. Feature extraction combining spectral noise reduction and cepstral histogram equalisation for robust ASR. In: *Proceedings of ICSLP 2002*, pp. 225–228, Denver, CO, USA.
- Steinbiss, V., Ney, H., Aubert, X., Besling, S., Dugast, C., Essen, U., Geller, D., Haeb-Umbach, R., Kneser, R., Meier, H.-G., Oerder, M., Tran, B.-H., 1995. The Philips research system for continuous-speech recognition. *Philips Journal of Research* 49, 317–352.
- Strik, H., Russel, A., van den Heuvel, H., Cucchiari, C., Boves, L., 1997. A spoken dialogue system for the Dutch public transport information service. *International Journal of Speech Technology* 2, 119–129.

- Tibrewala, S., Hermansky, H., 1997. Multi-band and adaptation approaches to robust speech recognition. In: Proceedings of Eurospeech 1997, pp. 2619–2622, Rhodes, Greece.
- Viikki, A., Laurila, K., 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication* 25, 133–147.
- Young, S., Jansen, J., Odell, J., Ollason, D., Woodland, P., 1997. *The HTK Book (for HTK Version 2.1)*. Cambridge University Press, Cambridge.