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

Author :Febe de Wet*, John de Veth, Loe Boves, Bert Cranen

Professor:陳嘉平
Reporter:葉佳璋

# Outline

- Introduction
- Time-domain noise reduction
- Experiment  result

# Introduction

- All start-of-the-art auto speech recognition systems are statistical pattern match machines. Their performance will deteriorate if there is a mismatch between the statistical properties of the training and testing.
- Different strategies have been developed to ensure that the statistical properties of training and test data are as  similar as possible.

  -condition raw speech signal

  -histogram normalization

# Introduction

- The impact of different transmission channel on logE features and MFCCs is similar: they cause a constant offset in the long-term mean of the feature track.

- However, the impact of additive noise on the shape of the overall distributions of logE features is very different from what is usually observed for cepstral features.

- It is reasonable to expect that non-linear transformations, which aim to equalize the mean and variance as well as the shape of the distribution, will have different effects on cepstral and logE energies parameters.

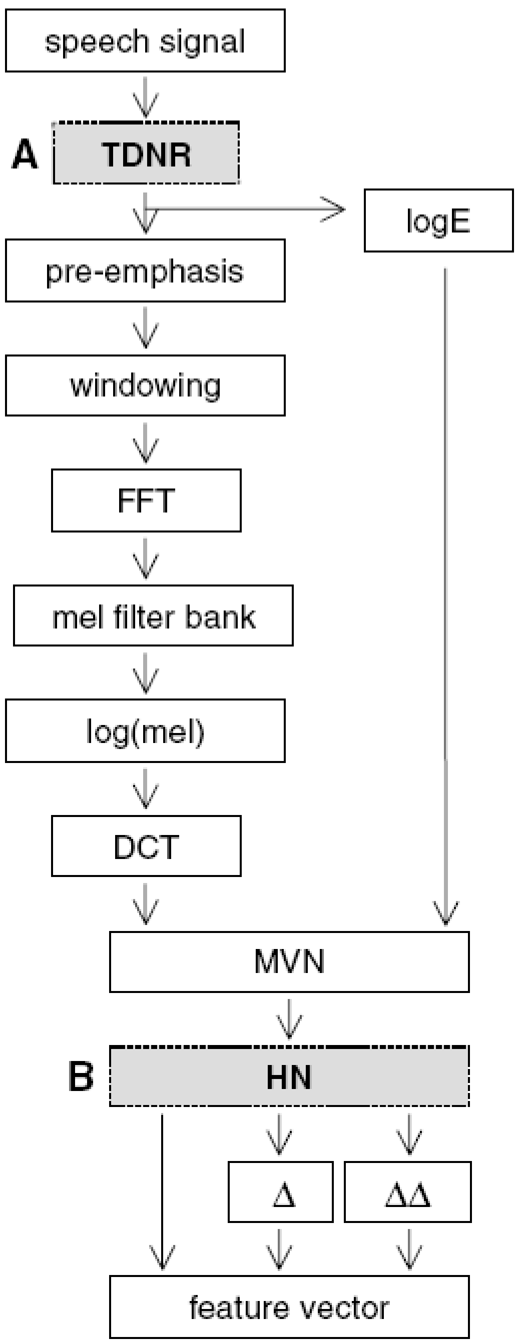


Fig. 2. Schematic overview of the feature extraction and mismatch reduction modules.

# Time-domain noise reduction

- 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.

  -If the VAD module classifies a frame as non-speech, it is used to update the estimate of 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.

# Time-domain noise reduction

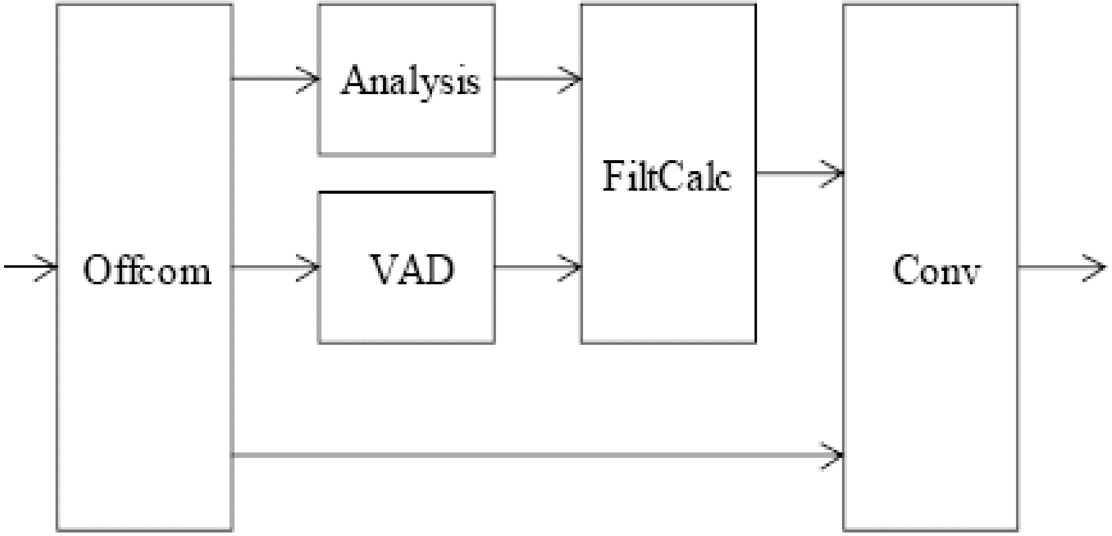
- These SNR estimate are subsequently used to derive the transfer function of a Wiener filter.

  -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 Wiener filter.

- Finally, the “clean”  signal is obtained by a convolving the original noisy signal with second-pass Wiener filter in the time domain.

# Time-domain noise reduction(TDNR)



# Speech and experimental set-up

- The speech data that was used in this study is a subset of the Aurora2 database.
- This study involves comparisons in terms of different feature types, different data transformations as well as different recognition task.
- It was decided not  to use the multi-condition training data because it would have complicated the experimental set-up considerably.
- Three test sets were defined for the Aourora2 task, i.e. sets A, B, C.

# Speech and experimental set-up

- The VIOS database was collected with an on-line version of a spoken dialog system that provides train time table in formation in the Netherlands.

  -The speech data was recorded over the public switched telephone net work.

  -None of the utterance used for training had a high background noise level

  -The noise data was collected in the hall of a train station in the Netherlands.

  -The noise was added to the original test data such that the resulting acoustic signals had SNRs of 0, 10 and 20 dB, respectively.

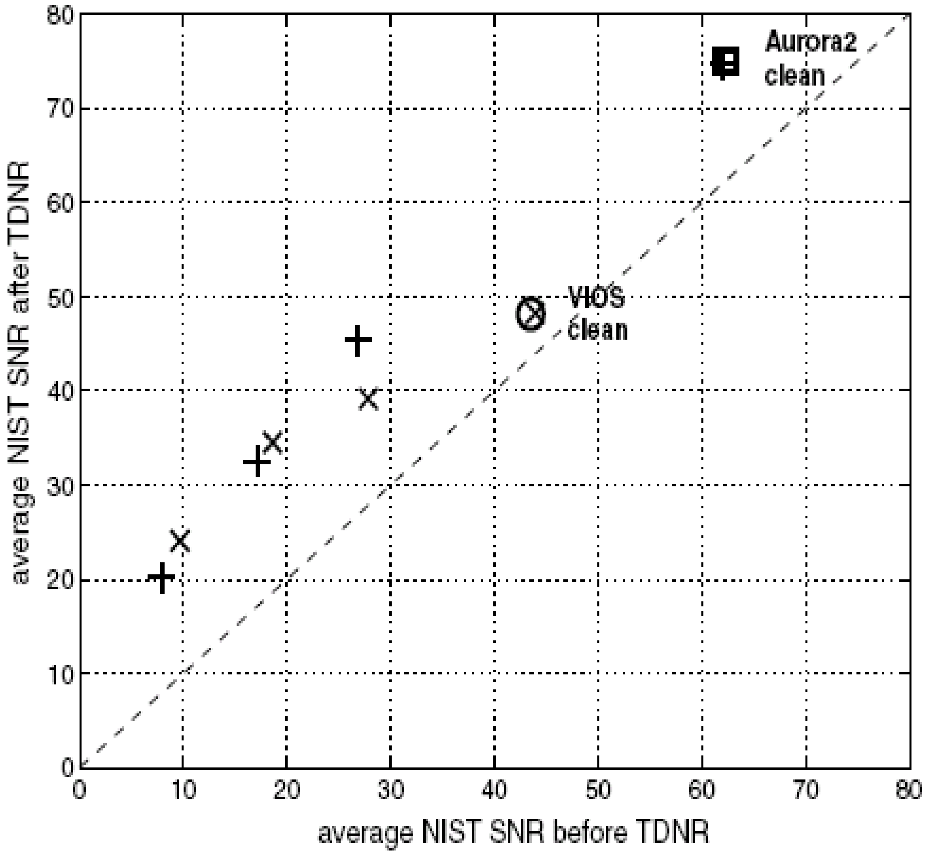


Fig. 1. Mean SNR of the clean and noisy (train station) Aurora2 (training (□), test (+)) and VIOS (training (○), test (×)) data before and after the application of TDNR.

- Three mismatch reduction experiment were carried out

  -In Experiment I, only block a in Fig. 2 (TDNR) was included in the acoustic pre-processing.

  -In Experiment II, only block b(HN) was included in the acoustic pre-processing.

  -In Experiment III, both block a and b were active.

- In each of the experiments, the mismatch reduction schemes were implemented in the following order:

  -not at all

  -only for the cepstral features and using the baseline logE

  -only for the energy and using the baseline MFCCs

  -for both MFCCs and the logE feature

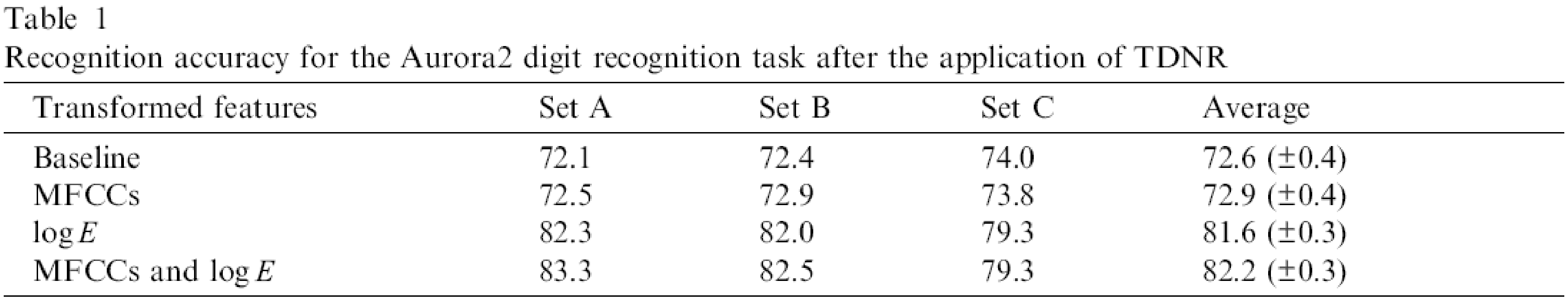


Table 1
Recognition accuracy for the Aurora2 digit recognition task after the application of TDNR

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

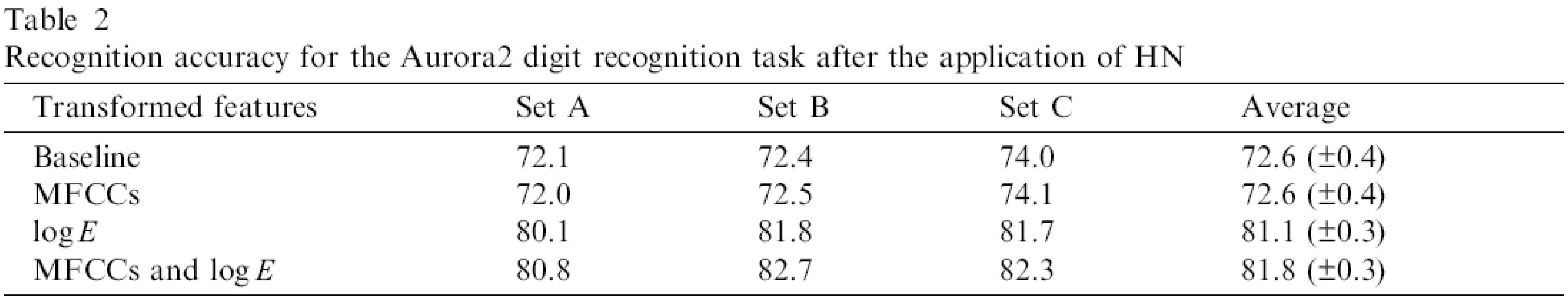


Table 2
Recognition accuracy for the Aurora2 digit recognition task after the application of HN

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

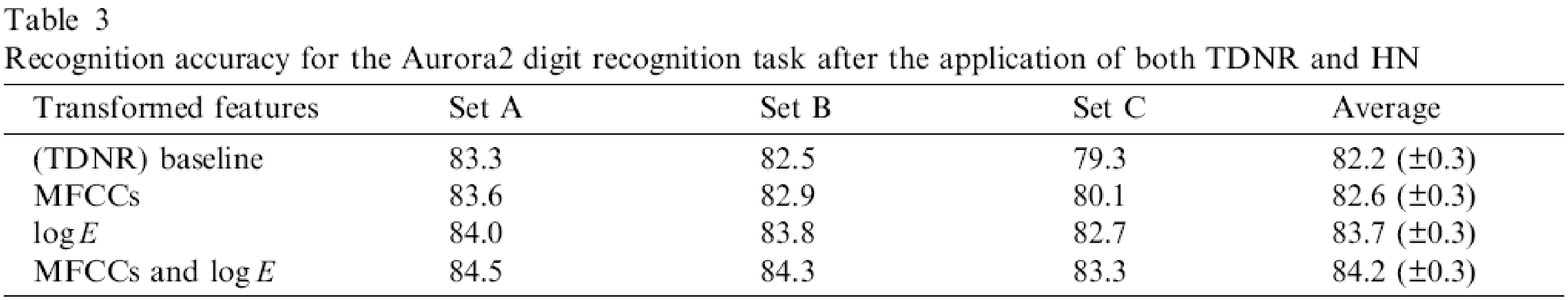


Table 3
Recognition accuracy for the Aurora2 digit recognition task after the application of both TDNR and HN

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

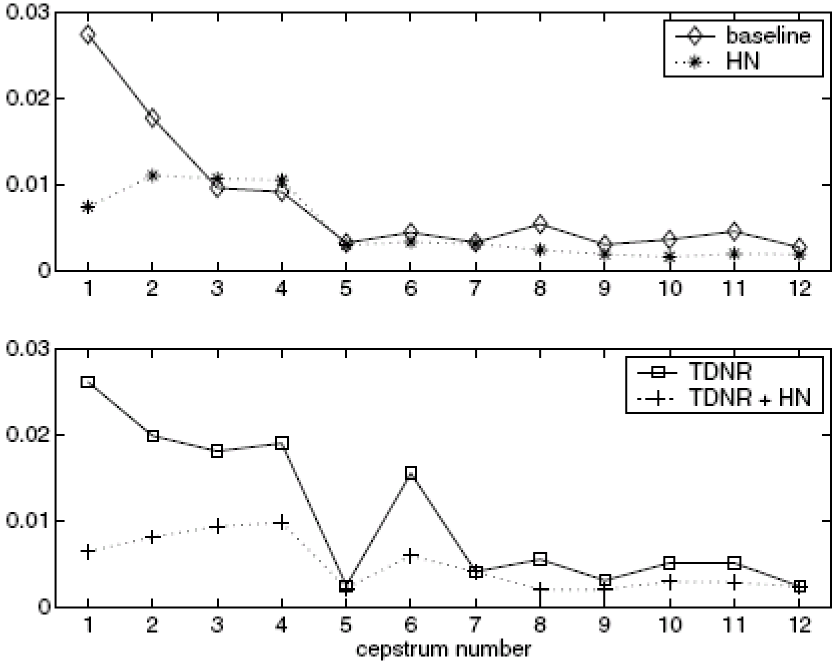


Fig. 4. Kullback divergence between the distributions of the training and test data for the 12 cepstral coefficients. Distances for the baseline condition and after the application of HN. (Top) Distances after TDNR and TDNR + HN have been applied (Bottom).

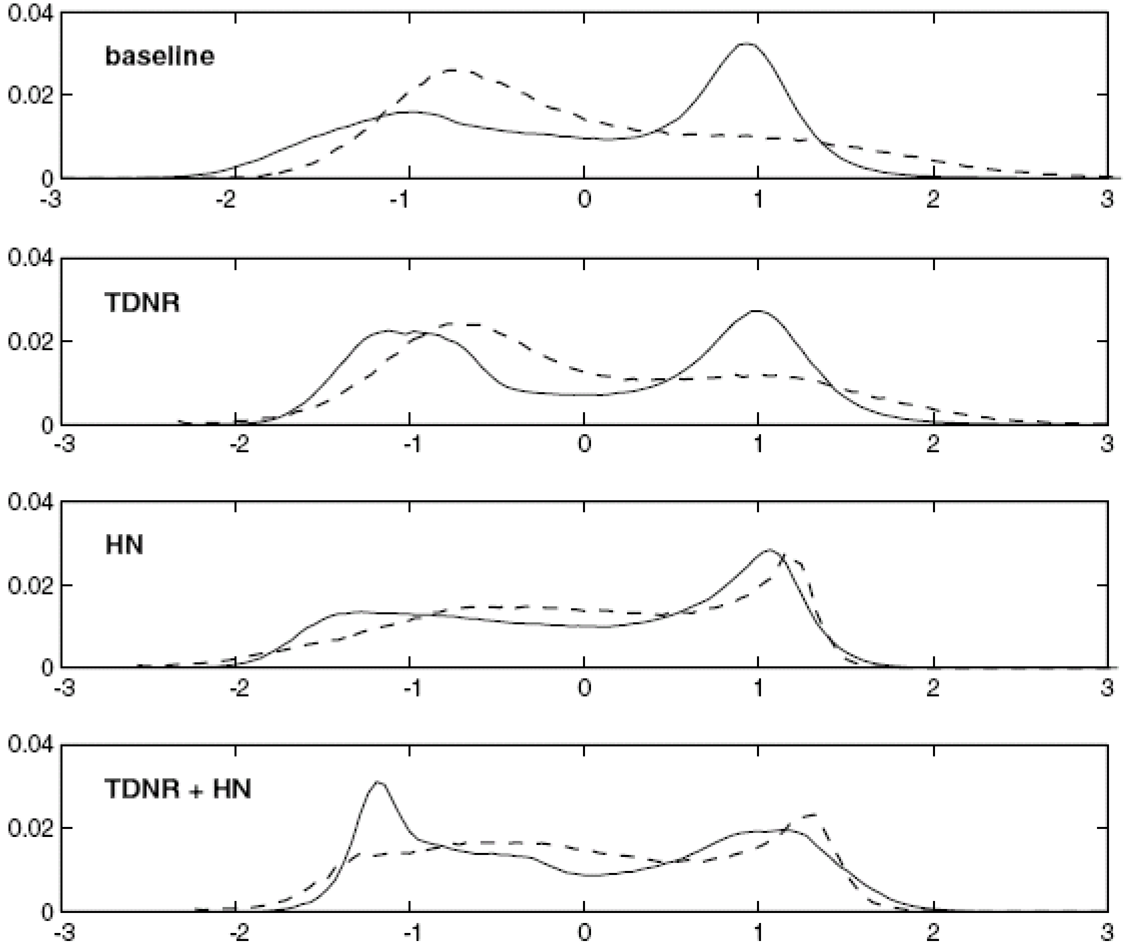


Fig. 5. Overall distribution of log$E$ derived from clean training data (solid line) and 0 dB SNR train station test data (dotted line) in the baseline condition, after the application of TDNR, after HN, and after TDNR + HN have been applied.

# Discussion

- The fact that reduced mismatch in the global distribution of the logE 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 logE feature may have a larger impact on recognition performance than MFCCs.

- In order to determine whether the observation made for the Aurora2 experiments generalize to a more complex task such as CSR.

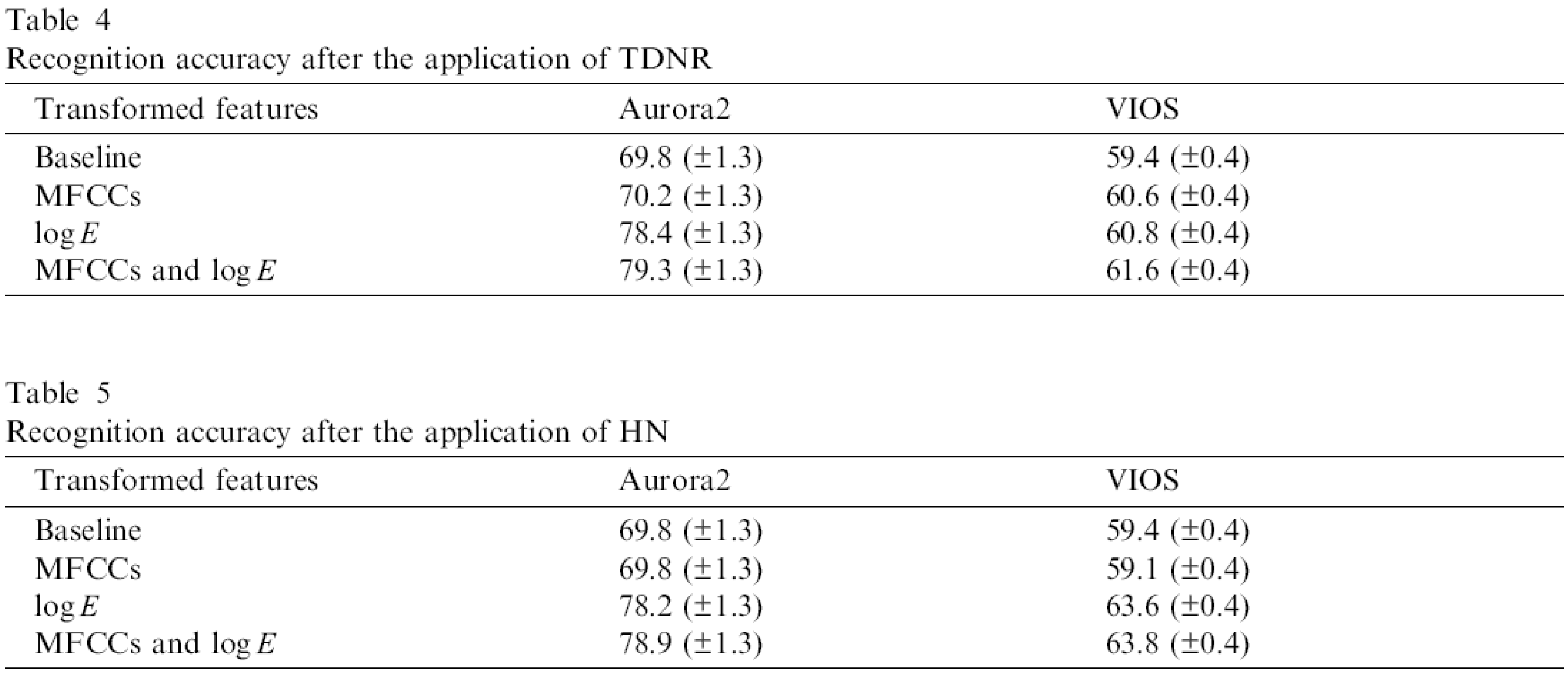


Table 4
Recognition accuracy after the application of TDNR

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

Table 5
Recognition accuracy after the application of HN

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

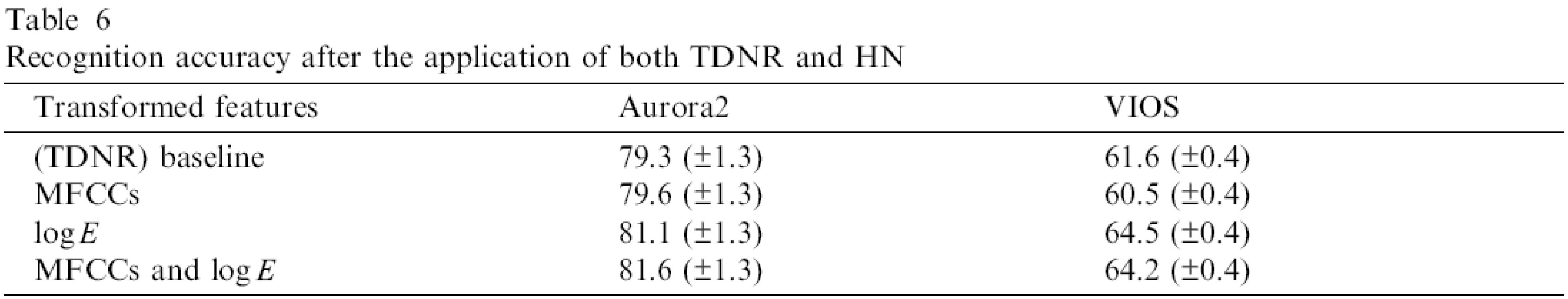


Table 6
Recognition accuracy after the application of both TDNR and HN

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

# Discussion

- A striking similarity between the Aurora2 and VIOS results is the significance of the logE feature: 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 logE features significantly improves recognition performance

- This observation seems to suggest that well-match logE feature are of primary importance for successful speech recognition, even in the more complex search space.

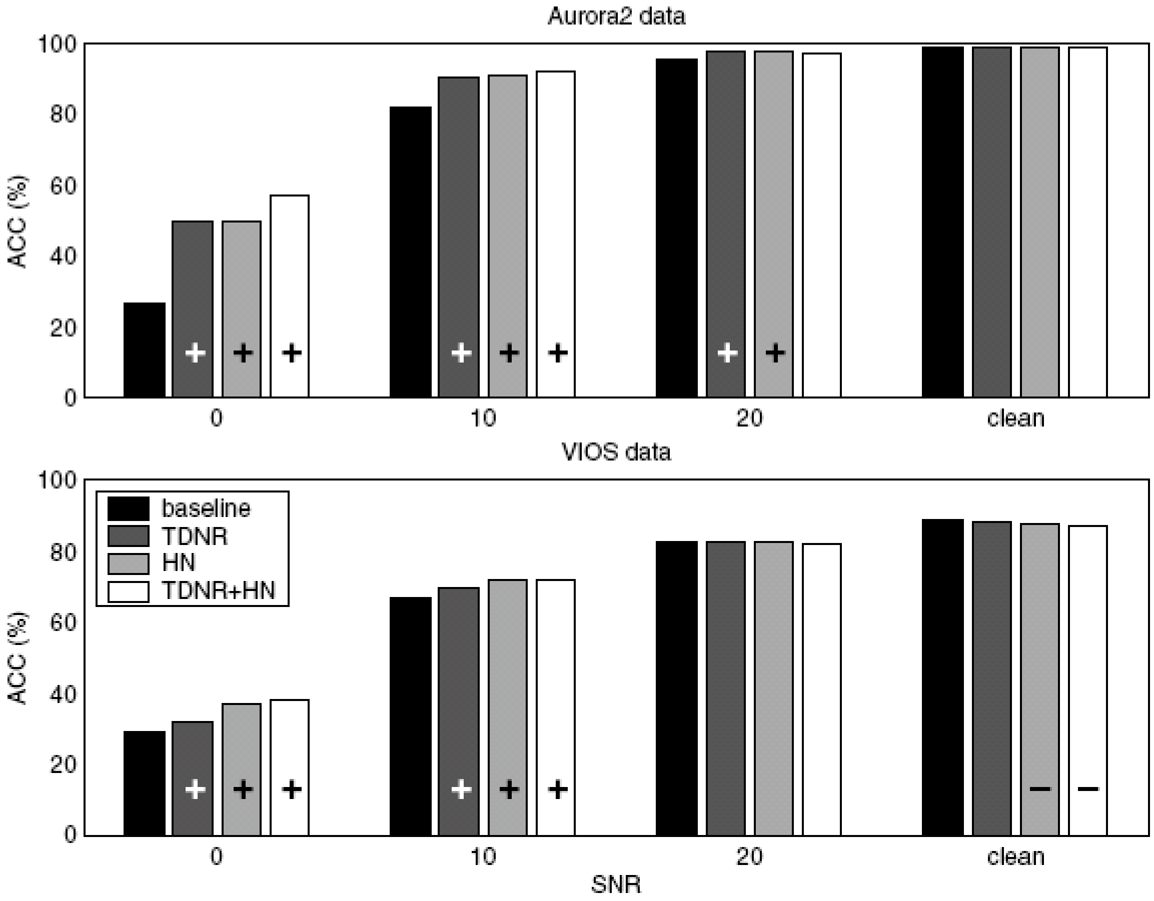


Fig. 6. Recognition accuracy as a function of SNR for the Aurora2 digit recognition task (top) and the VIOS CSR task (bottom). (significant increase (+), significant deterioration (–), no significant difference (empty bar) – relative to the baseline (filled bar)).