THE IMPACT OF SPECTRAL AND ENERGY MISMATCH ON THE AURORA2 DIGIT RECOGNITION TASK

Febe de Wet, Johan de Veth, Bert Cranen & Louis Boves

A^2RT, Department of Language & Speech
University of Nijmegen, The Netherlands
{f.dewet, j.deveth, b.cranen, l.boves}@let.kun.nl

ABSTRACT

Within the Aurora2 experimental framework, the aim of this study is to determine what the relative contributions of spectral shape and energy features are to the mismatch observed between clean training and noisy test data. In addition to measurements on the baseline Aurora2 system, recognition performance was also evaluated after the application of time domain noise reduction (TDNR) and histogram normalisation (HN) in the cepstral domain. The results indicate that, for the Aurora2 digit recognition task, TDNR, HN, as well as a combination of the two techniques achieve higher recognition rates by reducing the mismatch in the energy part of acoustic feature space. The corresponding mismatch reduction in the spectral shape features yields hardly any gain in recognition performance.

1. INTRODUCTION

The performance of automatic speech recognition (ASR) systems deteriorates substantially if there is a mismatch between the statistical distributions of the acoustic features derived from the training data and the corresponding distributions of the test data. Such a mismatch may have different causes, e.g. variations between speakers, differences between the acoustic properties of the training and test environment, differences in channel properties, etc. and will occur in any situation where the short term spectral properties of the test data - and consequently any features derived from the test data - differ from the corresponding short term spectral properties of the training data.

In this study the focus is on the statistical mismatch that occurs when models derived from clean training data are used to perform recognition in noisy acoustic environments. The presence of background noise in the test data usually results in a substantial difference between the statistical distributions of the training and test data. The impact of the background noise on the statistics of the test data may differ according to the type of background noise as well as the implementation details of the acoustic feature extraction process. The ultimate aim of robust ASR techniques is to ensure that training and test distributions are as similar as possible, no matter how the mismatch came about. In this investigation we used two methods to reduce training/test mismatch, i.e. time domain noise reduction based on Wiener filtering (TDNR) [2] and histogram normalization (HN) in the cepstral domain [3, 4, 5, 6].

TDNR is applied to the speech data itself and reduces training/test mismatch by processing noisy signals in such a way that they become more similar to clean signals. HN, on the other hand, compensates for the effects of the noise directly in the acoustic feature domain. HN involves a non-linear transformation of the acoustic feature vectors derived from the test data such that their statistical properties match those of the training data.

These two techniques to reduce training/test mismatch were used to investigate the relative importance of spectral shape and energy mismatch in the Aurora2 digit recognition task. Most ASR systems use acoustic feature vectors which describe acoustic data both in terms of the overall energy and the shape of the spectral envelope at frame level. These two types of features carry different information about the signal and it could therefore be expected that, under mismatched conditions, recognition performance might show a different sensitivity to mismatch in the energy features, on the one hand, and mismatch in the spectral shape parameters, on the other. In this study, we analysed the gain in recognition accuracy achieved by TDNR and HN in terms of mismatch reduction in the energy and spectral shape parameters. We used the log of the total energy in each frame (logE) as an energy feature and 12 mel-scaled cepstral coefficients (MFCCs) to describe the shape of the spectral envelope.

TDNR and HN are described in the next section. Section 3 gives an overview of the experimental design and set-up. A summary of the results is given in Section 4, followed by a discussion and conclusions in Sections 5 and 6.

2. MISMATCH REDUCTION TECHNIQUES

2.1. Time-domain noise reduction (TDNR)

The first mismatch reduction technique that was used in this study is the time-domain noise reduction scheme described in [2]. As a first processing step, offset compensation is applied to each utterance. 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 "denoised" 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

The experiments on the Aurora2 database were carried out within the framework of the SMADA project [1].

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.

2.2. Histogram normalisation (HN)

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 distributions 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 often assumed 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, should be derived such that:

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

In our implementation, 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 3^{rd} order spline function was used to approximate W_k. 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.

3. EXPERIMENTAL SET-UP

3.1. Speech data and recogniser

The speech data that was used in this study is a subset of the Aurora2 database, i.e. the clean condition training material and the three test sets. The Aurora2 database was derived from a subset of the TI-Digits database [7]. 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. In this study, we only report on the results obtained for the clean condition training, because it provides a more challenging noise reduction problem than the multi-condition experiments. Three test sets were defined for the Aurora2 task, i.e. sets A, B, and C. All three test sets are made up of a mixture of clean and noisy data. Sets A and B have the same channel properties as the training data but differ from each other in the types of noise they contain. In addition to the artificially added noise, the channel properties of the data in test set C also differs from that of the training data.

We used the reference recognition system that was developed for Aurora2 [7] in all the experiments described in this paper. The system is based on hidden Markov word models and implemented

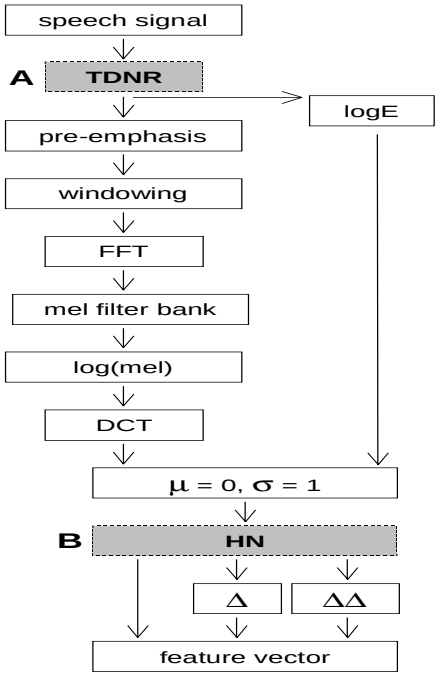

Fig. 1. Schematic overview of the feature extraction and mismatch reduction modules.

in HTK [8]. Each model has 16 states with 3 continuous density Gaussian mixtures 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 trained: one corresponding to silences at the beginning and the end of the utterances (3 states, 6 Gaussians per state) and one corresponding to silences between words (single state tied to the middle state of the 3-state silence model).

3.2. Acoustic pre-processing

Figure 1 gives an overview of the acoustic pre-processing procedure that was used to derive the spectral shape ($c_1 \ldots c_{12}$) and energy (logE) feature. The shaded blocks in the figure correspond 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 25ms 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 Discrete Cosine Transform. The log of the total energy (logE) was also calculated for each frame. In the experiments where TDNR was applied, logE was calculated from the "cleaned" signals.

In order to remove the effect of channel variation from the data we performed cepstral mean subtraction (at utterance level). Moreover, because we wanted HN to be independent of the range of the feature values, the MFCCs and logE values were normalised to have unit variance (at utterance level) according to the normal-

isation scheme described in [9]. 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 the features were transformed using HN, the first and second order time derivatives were derived after applying HN.

3.3. Mismatch reduction experiments

Three mismatch reduction experiments were carried out. In the first experiment (Experiment A), only block A in Figure 1 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 the second experiment (Experiment B), only block B was included in the acoustic pre-processing, i.e. HN was applied to the MFCCs and logE *after* feature extraction. In the third experiment (Experiment C) 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 spectral shape features (MFCCs) 3) only for the energy feature (logE) 4) for both the spectral shape and the energy 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

The results in this section are defined in terms of recognition accuracy, i.e. $\frac{N-S-D-I}{N} \times 100\%$, 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. The values shown in the tables below 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 $\times$ SetA + 0.4 $\times$ setB + 0.2 $\times$ setC.

4.1. Experiment A: TDNR

Table 1. Recognition accuracy after the application of TDNR.

Transformed features	Set A	Set B	Set C	**Average**
baseline	72.10	72.41	74.02	**72.61**
MFCCs	72.52	72.93	73.77	**72.93**
logE	82.34	81.98	79.31	**81.59**
both	83.32	82.52	79.28	**82.19**

The results for Experiment A are summarised in Table 1. The values in the table show that, calculating only the MFCCs from the data after applying TDNR, yields a marginal increase in recognition accuracy. The results also show that reducing the mismatch in logE accounts for most of the total gain in recognition rate, i.e. when only logE is calculated from the "cleaned up" data (i.e. after the application of TDNR), the recognition rates are only slightly inferior to those obtained when both the MFCCs and logE are calculated from the "cleaned up" data.

4.2. Experiment B: HN

Table 2 gives an overview of the results obtained in Experiment B. According to the values in Table 2, the recognition performance obtained when HN is applied only to the MFCCs does not differ much from the baseline. However, applying HN on logE yields a marked increase in recognition rate. When both the MFCCs and logE are transformed, the largest part of the total gain can be attributed to the transformation applied to logE. This trend in the results was also observed in experiment A. However, the results for test sets A and C in the last two rows of Table 2 differ substantially from those that were measured in Experiment A: in Table 2 the mean recognition accuracy for test set A is almost 3% lower and the mean recognition accuracy for test set C is almost 3% higher than in Table 1. An analysis of the individual test sets revealed that this difference can 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 seem to 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.

Table 2. Recognition accuracy after the application of HN.

Transformed features	Set A	Set B	Set C	**Average**
baseline	72.10	72.41	74.02	**72.61**
MFCCs	72.03	72.51	74.08	**72.63**
logE	80.11	81.80	81.65	**81.09**
both	80.80	82.68	82.25	**81.84**

4.3. Experiment C: TDNR & HN

The recognition accuracies that were measured in Experiment C 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. Transforming only the MFCCs yields a small gain in recognition performance. As was observed in the previous experiments, reducing the mismatch in logE results in the biggest gain in recognition accuracy. However, the best results are obtained when both the MFCCs and logE are transformed.

Table 3. Recognition accuracy after the application of both TDNR and HN.

Transformed features	Set A	Set B	Set C	**Average**
(TDNR) baseline	83.32	82.52	79.28	**82.19**
MFCCs	83.63	82.87	80.13	**82.63**
logE	84.02	83.82	82.74	**83.68**
both	84.46	84.26	83.33	**84.16**

5. DISCUSSION

As was mentioned in Section 1, acoustic mismatch between training and test data leads to mismatch in both the energy and the spectral shape components of acoustic feature space. The results that were presented in the previous section show that, for the Aurora2 digit recognition task, both TDNR and HN improve recognition rate by reducing the mismatch in energy features rather than their cepstral counterparts: in both experiments A and B more than 90% of the total gain in recognition performance could be attributed to a reduction of the mismatch in logE. The effect is less prominent but still clearly visible in the results of Experiment C.

Given the difference in performance for TDNR and HN on test sets A and C, the fact that a combination of the two methods - as was applied in Experiment C - gives the best overall results does not really come as a surprise. The results suggest that, after the application of TDNR, there is still a residual mismatch between the distributions of the training and test features. This mismatch is observed most clearly in the distribution of the logE features and can be compensated for by applying HN in combination with TDNR.

In [10] the authors reported a large gain in recognition rate for the Aurora2 task if spectral subtraction was applied in combination with HN in the cepstral domain. They attributed this result to the ability of HN to compensate for the linear channel distortion and the residual non-linear distortions remaining after spectral noise reduction. In contrast, the results reported in [5] show almost no improvement when HN is applied in the cepstral domain. However, in that study, the authors also applied cepstral mean and variance normalisation, while channel normalisation was not performed prior to HN in [10]. The results reported in the current study 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 system. A substantial increase in recognition rate is only observed when HN is applied to logE. These results, together with the results reported in [5] suggest that, after cepstral mean subtraction and variance normalisation, a non-linear transformation of cepstral coefficients does not improve recognition performance, whereas a non-linear transformation of logE does lead to higher recognition accuracies. These observations indicate that the gain reported in [10] may probably be attributed to channel compensation (in the cepstral domain) and that the non-linear distortions remaining after spectral noise reduction are probably limited to the energy features.

6. CONCLUSIONS

The results of this study show that, for the Aurora2 digit recognition task, TDNR, HN, as well as a combination of the two techniques achieve higher recognition rates by reducing the mismatch in the energy part of acoustic feature space. The corresponding mismatch reduction in the spectral shape features yields hardly any gain in recognition performance. The results also show that it is possible to achieve almost the same improvement in recognition accuracy by applying a relatively simple HN transformation in the acoustic feature domain than by applying noise reduction to the signals in the time domain. Moreover, it has also been shown that, after the application of TDNR, the application of HN on MFCCs yields very little additional gain in recognition rate. This observation seems to suggest that, if there is still a non-linear mismatch between the distributions of the training and test features after the application of TDNR, it is most probably limited to the distributions of the logE features.

These observations may be a result of the fact that the mismatch reduction techniques that were used in this study are better at reducing mismatch in energy features (logE) than in spectral shape features (MFCCs). We are currently investigating various ways to quantify the degree and type of mismatch between different data sets in terms of their energy and spectral shape features in order to verify this supposition.

The fact that reduced mismatch in logE leads to such large improvements in recognition performance may also be an artefact of the experimental set-up that has been chosen for Aurora2. Because of the low complexity of the task, the logE feature may have a larger impact on recognition performance than the MFCCs. This may not be the case for a more complex task such as continuous speech recognition (CSR). In the near future we will repeat these experiments on continuous speech in order to determine to what extent the observations that were made for the Aurora2 task generalise to CSR.

7. REFERENCES

[1] L. Boves, D. Jouvet, J. Sienel, R. de Mori, F. Bechet, L. Fissore, and P. Laface, "ASR for automatic directory assistance: the SMADA project," in *Proceedings of ASR 2000*, Paris, France, 2000, pp. 249–254.

[2] B. Noe, J. Sienel, D. Jouvet, L. Mauuary, L. Boves, J. de Veth, and F. de Wet, "Noise reduction for noise robust feature extraction for distributed speech recognition," in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 433–436.

[3] S. Dharanipragada and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition," in *Proceedings of ICSLP 2000*, Beijing, China, 2000, pp. 556–559.

[4] F. Hilger and H. Ney, "Quantile based histogram equalisation," in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 2001.

[5] S. Molau, M. Pitz, and H. Ney, "Histogram normalisation in the acoustic feature space," in *Proceedings of ASRU 2001*, Madonna di Campiglio, Trento, Italy, 2001.

[6] J.C. Segura, M.C. Benitez, A. de la Torre, S. Dupont, and A. Rubio, "VTS residual noise compensation," in *Proceedings of ICASSP 2002*, Orlando, USA, 2002.

[7] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ASR 2000*, Paris, France, 2000, pp. 181–188.

[8] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, *The HTK Book (for HTK Version 2.1)*, Cambridge University, Cambridge, UK, 1997.

[9] A. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[10] J.C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Feature extraction combining spectral noise reduction and cepstral histogram equalisation for robust ASR," in *Proceedings of ICSLP 2002*, Denver, USA, 2002.