Automatic Speech Recognition in Adverse Acoustic Conditions

Febe de Wet, Johan de Veth, Bert Cranen & Loe Boves
A^2RT, Department of Language & Speech, University of Nijmegen
{F.de.Wet, J.deVeth, Cranen, L.Boves}@let.kun.nl

Abstract

Automatic Speech Recognition (ASR) technology has reached maturity to the extent that it can be used successfully in various applications. However, it is by no means the "solved problem" that some marketing campaigns are promoting it to be. One of the biggest challenges that operational ASR systems are faced with, is to maintain recognition performance in adverse acoustic conditions. The training procedures of most ASR systems yield recognisers with a relatively rigid image of the world: Only those acoustic variations that actually occurred in the training data are accounted for. Since training data is usually clean (in the sense that care is taken to avoid noisy recording environments, channel noise, etc.), noise sources which are present when the system is operational result in a mismatch between the training and the test conditions. Such a mismatch may reduce recognition performance quite significantly The aim of this research is to determine the extent to which the robustness of ASR systems against mismatched training and test conditions may be increased using acoustic backing-off as an implementation of Missing Feature Theory

Introduction

Automatic Speech Recognition (ASR) technology has reached maturity to the extent that it can be used successfully in various applications. However, guiding experimental systems on their way out of the laboratory has proven to be all but a trivial task. ASR systems which are operational under "real world" conditions are faced with a number of daunting challenges. One of these lies in the fact that they are often required to perform in adverse acoustic conditions. But what exactly are adverse acoustic conditions? Given the basic principles of ASR, the answer may be phrased as follows:

Generally speaking, ASR systems are able to recognise speech sounds automatically by comparing their acoustic characteristics with those of statistical models that were constructed during training. The basic idea of the training phase is to build general speech sound models which describe the statistical properties of all the relevant acoustic characteristics of speech that may occur in practice. However, the training procedures of most ASR systems yield recognisers with a relatively rigid image of the world: Only those acoustic variations that actually occurred in the training data are accounted for. Since training data is usually clean (in the sense that care is taken to avoid noisy recording environments, channel noise, etc.), noise sources which are present when the system is operational result in a mismatch between the training and the test conditions. If such a mismatch leads to significantly different statistical distributions of the acoustic characteristics, the ASR is said to be operating in adverse acoustic conditions.

It may also be argued that the statistical models can be trained on noisy data in order to alleviate possible acoustic mismatches between training and test conditions. Such a procedure would require training data that is representative of all possible acoustic conditions in which the system may be used. Even if enough training material could be gathered to describe these environments adequately, it would most probably result in statistical models whose discriminative ability does not go beyond a distinction between noise and speech. This line of reasoning does present a rather extreme view on the matter, but it serves to illustrate how difficult the acoustic modelling of speech is if the acoustic characteristics of the environment within which the speech is produced is unknown.

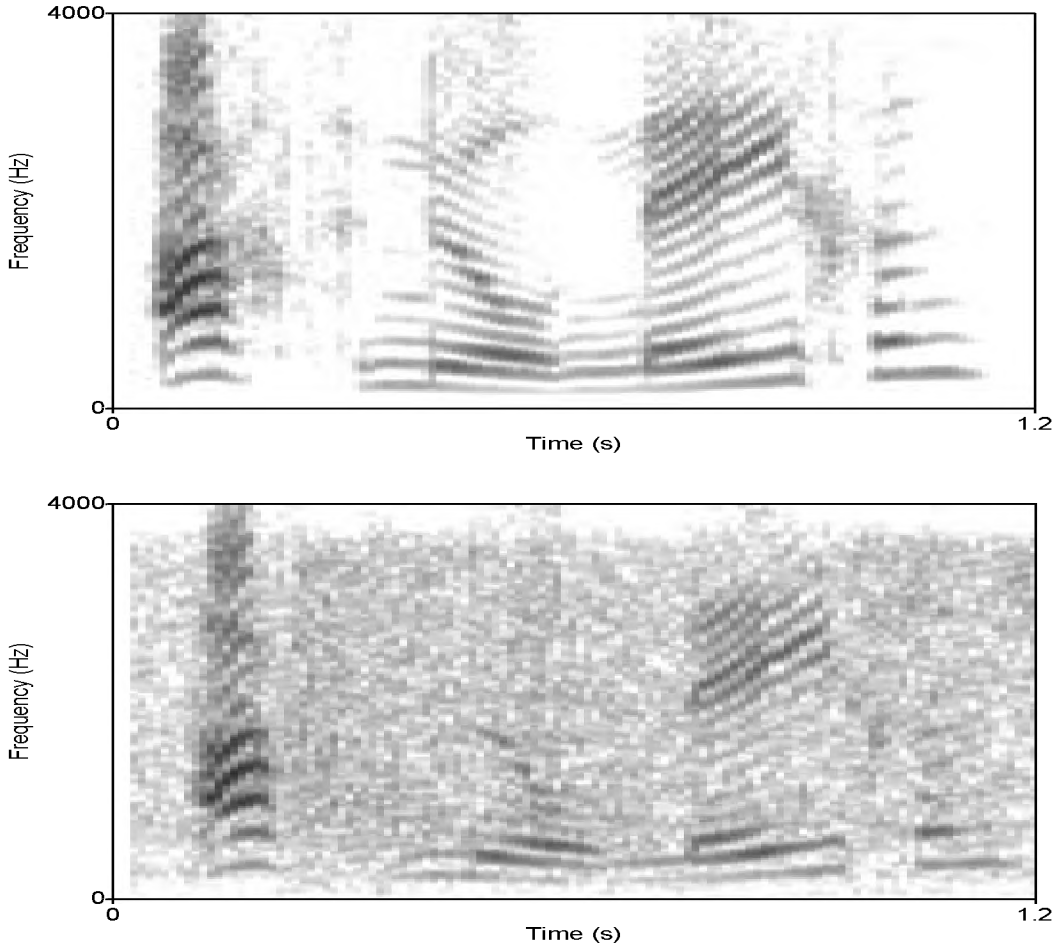

Figure 1: ***Spectrogram representation of the utterance "acht nul negen" in clean (top) and noisy (bottom) conditions.***

A typical example of the effect of adverse acoustic conditions on a speech signal is illustrated in Figure 1. The top part of the figure shows the spectrogram of the utterance "acht nul negen", i.e. Dutch for "eight zero nine". The speech was recorded over a telephone line under clean conditions and would typically be used as an example of clean speech during training. The bottom part of the figure represents the same utterance produced in adverse acoustic conditions. (Factory noise from the NOISEX CD (Noisex, 1990) was artificially added to the speech signal to simulate adverse acoustic conditions.) A comparison of the two spectrograms clearly reveals that there is a significant degree of mismatch between the two signals. If the noisy

signal was to be presented to an ASR system trained only on clean data, the mismatch would most certainly have a detrimental effect on system performance. Many words would be recognised incorrectly because the spectral properties represented by the models in the ASR system do not match those contained in the signal. Ideally ASR systems should be able to maintain recognition performance, no matter what the acoustic conditions are. Recently, it was suggested that Missing Feature Theory (MFT) can be used to improve the robustness of ASR under adverse acoustic conditions (Cooke et al., 1996; Lippmann & Carlson, 1997; Morris et al., 1998). By using only the reliable parts of the acoustic information and disregarding unreliable acoustic features, recognition performance can almost be maintained at the level for undisturbed conditions.

In the recognition engines of standard ASR systems, feature distributions are often modelled by means of Gaussian probability density functions. During recognition a local distance function is used to determine how likely it is that an observation value belongs to the model of a given sound. The local distance function, calculated as the negative of the natural logarithm of the Gaussian probability function, is a quadratic function. It is rather unlikely that the tails of a Gaussian distribution are reliable estimators of the less frequently occurring feature values. As a consequence, it might not be such a good idea to define the contribution to the local distance function used during dynamic programming as a quadratic function over the entire feature value range.

In de Veth et al. (1998) and de Veth et al. (1999) it was proposed to model feature value observations by means of two distributions: one obtained from the training data and another, uniform distribution which represents all feature values not seen during training. In contrast to the way the *conventional local distance function* is calculated, i.e. as a quadratic function over the entire feature value range, this computation interpolates between the two distributions. The weight assigned to either distribution can be varied so as to increase or decrease the contribution of the unseen values. This strategy was called *acoustic backing-off* and it was shown that it can be considered as an implementation of MFT which (1) is suited to be used in a conventional ASR system, (2) in principle allows one to use any feature representation as long as at least part of the acoustic feature vector is undisturbed, (3) contrary to the approach suggested in Lippmann & Carlson (1997) does not

require prior information about the corrupted features and (4) does not rely on an explicit detection mechanism for identifying disturbed feature vector elements as opposed to the approaches suggested in Dupont et al. (1997) and Tibrewala & Hermansky (1997).

However, the application of MFT is not as straightforward as it might seem, since there appears to be an interaction between the features that are severely affected (and therefore unusable during recognition) and the signal pre-processing steps associated with typical ASR systems (de Veth et al., 1999). Raw spectral representations, such as those illustrated in Figure 1, are rarely used to build statistical models for ASR systems. Instead normalising (e.g. gain normalisation, channel normalisation) and orthogonalising (e.g. Discrete Cosine Transform (DCT), Linear Discriminant Analysis (LDA)) transforms are widely used in state-of-the-art ASR systems to generate the features that are used in the actual recogniser. The main reason for using normalisation transforms is to remove non-speech biases that are introduced to the signal by e.g. transmission channel characteristics and recording equipment. If these signal components are removed before calculating the features, their characteristics will not influence the model statistics, i.e. the aim of normalisation is to ensure that model means and variances are based on speech information only. Orthogonalisation is generally applied to remove the correlation between raw spectral features so that a full-covariance matrix can be replaced by a diagonal covariance matrix. A diagonal matrix is preferred because its elements can be estimated reliably with less data. For clean speech data, normalisation and orthogonalisation transforms generally improve recognition performance significantly.

In this article the aim is to show why the simultaneous application of MFT to reduce the detrimental effect of adverse acoustic conditions on the one hand, and normalisation and orthogonalisation transforms on the other hand, may become undesirable. An intuitive understanding of the reasons behind this incompatibility may be obtained by considering the following reasoning: The basic pre-supposition in MFT is that a feature vector can be considered to consist of a part which is virtually unaffected and another part which contains distorted features. As long as the loss of information about the speech signal represented by the disturbed features is relatively small, MFT predicts that recognition performance can be maintained at a level which is comparable to the undisturbed case, simply by discarding

the disturbed features. However, a complication arises when the raw incoming features are first transformed by means of an algorithm which uses all feature vector elements to calculate a transformed vector. In this case, the misleading information due to the disturbances which are present in a restricted number of raw features, will be smeared out over the entire normalised/orthogonalised vector. If this happens, there is little hope that MFT can effectively help in recovering from the disturbances. Results of previous studies have shown that every possible effort should be taken to minimise the dispersion of disturbances (de Veth et al., 1999b; de Veth et al., 1999a). The investigations up until now mainly focus on the effects of within-vector smearing, but it is to be expected that the same holds true for the time (across vector) dimension.

In the rest of the paper, it will be assumed that the incoming speech is represented as a set of mel frequency log energy coefficients (MFLECs). To distinguish these input vectors from the feature vectors that result from preprocessing, i.e. those which are actually used for recognition, the MFLECs will be called *raw features*. The vector elements that result *after* pre-processing will be referred to as *feature values*.

The statistical models used during experimentation were based on three different feature representations, i.e.:

- mel-frequency cepstral coefficients (MFCCs),
- sub-band mel-frequency cepstral coefficients (SB-MFCCs) (Okawa et al., 1998) and
- within vector filtered mel-frequency log-energy coefficients (WVF-MFLECs) (Nadeu et al., 1995)

Details about these feature representations will be given in the section on acoustic features. For the moment it suffices to note that the first representation (MFCCs) is calculated from the entire vector of raw input features. As a consequence, any distortion in the raw input features is dispersed over all feature values that are used during recognition. In the rest of this paper this property of MFCCs will be referred to as *full smearing*, i.e. spectrally local distortions are "smeared" over the whole feature vector. The

other two representations (SB-MFCCs and WVF-MFLECs) are designed so that distortions which are present in part of the raw input feature vector do not necessarily spread over the entire feature vector that results after pre-processing. In other words, given the type of distortion applied, these representations guarantee that part of the feature vector remains unaffected. The SB-MFCCs and WVF-MFLECs feature representations will therefore be referred to as *partially smearing*.

In the next three sections, the experimental set-up of this investigation is introduced. Following on the description of the experiment, the recognition performance for the three different types of acoustic features is compared. Recognition performance was evaluated with clean and disturbed data for each of the three acoustic representation techniques, with and without applying MFT in the form of acoustic backing-off. The conclusions drawn from the experimental results are presented in the last section of the paper.

Speech Material

The speech material for the experiments was taken from the Dutch POLYPHONE corpus (den Os et al., 1995). Speech was recorded over the public switched telephone network in the Netherlands. Among other things, the speakers were asked to read several connected digit strings. The number of digits in each string varied between 3 and 16. A set of 1997 strings (16582 digits) was reserved for training. Care was taken so as to balance the training material with respect to sex, region (an equal number of speakers from each of the 12 provinces in the Netherlands) and the number of tokens per digit. 504 digit string utterances (4300 digits) were used for cross-validation during training cf. (de Veth & Boves, 1998). The system was evaluated with an independent test set of 1008 test utterances (8300 digits). The cross-validation and independent test sets were balanced according to the same criteria as the training material. None of the utterances used for training or testing had a high background noise level.

Acoustic Features

Three different types of acoustic features were used during experimentation: mel-frequency cepstral coefficients (MFCCs), sub-band mel-frequency cepstral coefficients (SB-MFCCs) and within-vector filtered mel-frequency log-energy coefficients (WVF-MFLECs). The features were calculated as follows:

The speech signals in the POLYPHONE corpus were recorded from a primary rate ISDN telephone connection and stored in A-law format. These were first converted to linear PCM format. A 25 ms Hamming window shifted with 10 ms steps and a pre-emphasis factor of 0.98 were then applied to the linear data. The data was subsequently converted to the frequency domain by applying the fast Fourier transform. In the frequency domain 16 filtered band energy values were calculated. The filter bands were triangularly shaped and uniformly distributed on a mel-frequency scale (covering 0-2143.6 mel; this corresponds to the linear range of 0-4000 Hz). Finally the log of each filter bank output was calculated to yield 16 mel-frequency log-energy coefficients (MFLECs) for each 25 ms interval of the speech signal. In addition to the 16 MFLECs, the log-energy for each frame was also computed. These signal processing steps were performed using HTK2.1 (Young et al., 1995).

Twelve MFCCs were computed from the raw MFLECs using the DCT. Cepstral mean subtraction (CMS) was then applied as a channel normalisation (CN) technique. The off-line version of this CN technique was used, i.e. the cepstral mean was computed using the whole utterance. The time derivatives of the MFCCs were computed and added to the 12 channel normalised feature values. The log-energy and delta log-energy values of each frame were also included, i.e. 26-dimensional acoustic feature vectors were obtained in this manner.

Twelve SB-MFCCs were derived from the raw MFLECs by computing a set of 6 MFCCs from the first 8 MFLEC values (covering 0 - 1218 Hz) and another set of 6 MFCCs from the second 8 MFLECs (covering 1015 - 4000 Hz). The rest of the calculations are exactly the same as those used to obtain MFCCs, i.e. subtract the mean computed over the whole utterance for CN

and compute the deltas. Together with log-energy and delta log-energy this procedure yielded 26-dimensional feature vectors.

For each frame coefficients 2 - 15 of the WVF-MFLECs were calculated by applying the algorithm:

$$c_i = MFLEC_{i+1} - MFLEC_{i-1} \quad (1)$$

to the corresponding frame of raw MFLECs. The values for coefficients 1 and 16 were copied from the raw MFLECs vector. After this filter and copy operation, the mean value computed over the whole utterance was subtracted as a form of CN and the delta features were computed. The static and delta WVF-MFLECs were combined with log-energy and delta log-energy to create 34-dimensional feature vectors.

Statistical Models

Hidden Markov models (HMMs) were used to describe the statistics of the speech sounds. A phone-based system was used, i.e. the basic speech sounds that were to be recognised were phones. The ten Dutch digit words were described with 18 phone models. Three additional models were used to capture the statistical properties of the silence, background noise and out-of-vocabulary speech in the recordings of the POLYPHONE database. Each phone unit was represented as a left-to-right HMM consisting of three states with the emission probability density function (pdf) of each state in the form of a single Gaussian pdf. Only self-loops and transitions to the next state were allowed. For these models the total number of different states was 63 (54 for the phones plus 9 for the noise models). HTK2.1 was used for training and testing the HMMs (Young et al., 1995). Training was done according to the cross-validation scheme described in de Veth & Boves (1998). The more complex models were obtained through subsequent mixture splitting. The single Gaussian pdf was split four times, resulting in different recognition systems with 2, 4, 8 and 16 Gaussians per state (containing respectively 126, 252, 504 and 1008 Gaussians in total). All HMMs were implemented using diagonal covariance matrices and each model set was trained only once, using clean speech data, i.e. undisturbed features.

The recognition syntax used during cross-validation and testing allowed for digit strings, varying in length from 3 to 16 digits, to be recognised.

Simulating Adverse Acoustic Conditions

Ideally speaking, the ultimate aim of this project is to find an acoustic representation technique which is immune to any arbitrary type of noise e.g. broad band, non-stationary, etc. However, given the knowledge that is currently available on this subject and the restrictions on the modelling framework, this problem is far too complex to handle. It was therefore decided to start the investigation with a simplified "noise problem" in order to gain insight into the way the different acoustic representations are affected by different kinds of noise (as measured by a degradation in recognition performance). In previous experiments band limited, stationary noise was added to the speech signals such that the resulting signal to noise ratios (SNRs) were 20, 10 and 5 dBA[1] (de Veth et al., 1999a).

Since there seems to be no qualitative difference in system behaviour at the various SNRs, only the 10 dBA SNR data is used in this paper. It represents a noise condition which is far from ideal but also does not cause recognition to fail completely. The effect of different frequency ranges of band limited noise on system performance is subsequently investigated at this level of distortion. For each experiment the band limited noise signals were obtained by filtering Gaussian white noise with a fifth order elliptical filter. The cut-off frequencies of the band-pass filters were chosen such that approximately one quarter of the resulting MFLECs were contaminated. Three different frequency ranges were investigated. These will be referred to as *low*, *middle* and *high*. For the low frequency range ($F_{low} = 395$ Hz, $F_{high} = 880$ Hz) the high cut-off frequency was chosen such that the noise distortions were limited to the first sub-band in the case of the SB-MFCC feature representation. The mid-frequency range ($F_{low} = 833$ Hz, $F_{high} = 1446$ Hz) was chosen in the middle of the mel frequency scale so that both sub-bands of the SB-MFCC features would be affected. The low

[1]Both the speech and noise energy levels were weighted according to the A-scale (Hassall & Zaveri, 1979).

cut-off frequency for the high range ($F_{low} = 1446$ Hz, $F_{high} = 2303$ Hz) ensured that only the second sub-band of the SB-MFCC features would be affected by the additive noise.

Results and Discussion

During training the best HMMs were determined by means of the cross-validation technique described in de Veth & Boves (1998). In order to determine a proper reference system for each feature representation, a word error rate (WER) was computed for the test set at 1, 2, 4, 8 and 16 Gaussians per state with WER defined as:

$$WER = \frac{S + D + I}{N} \times 100\%, \tag{2}$$

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 best results were obtained at 16 Gaussians per state where WER values varied between 2.4% (WVF-MFLECs) and 3.3% (SB-MFCCs). All subsequent recognition experiments were therefore performed using HMMs with 16 Gaussian mixture components per state. The results obtained for the three different feature sets under investigation at this working point, are shown in column 1 of Figure 2. As can be seen, the WER values of MFCCs and SB-MFCCs do not show substantial differences. However, the WVF-MFLECs representation yielded significantly better results. This finding is in good agreement with observations reported in Nadeu et al. (1995).

Using the low, mid and high range noise distortions, system performance was evaluated using a recognition system based on a *conventional local distance function*. The results are summarised in Figure 2. According to the results in columns 2 and 4 of Figure 2, the WVF-MFLECs and SB-MFCCs systems perform significantly better than their MFCC counterpart for the low and high noise conditions. They achieve WERs between 34.1% and 38.1% while the corresponding results for the MFCC system are between 46.7% and 64.7% for the two respective cases. These results correspond with previous observations that recognition performance suffers most for

feature representations that smear spectrally local distortions over *all* feature vector components (de Veth et al., 1999b). This observation shows that limiting the dispersion of the distortions that are present in the raw input features to only a sub-set of the feature vector components obtained after pre-processing helps to reduce the detrimental effect of the distortions, even with a conventional local distance function, i.e. without acoustic backing-off.

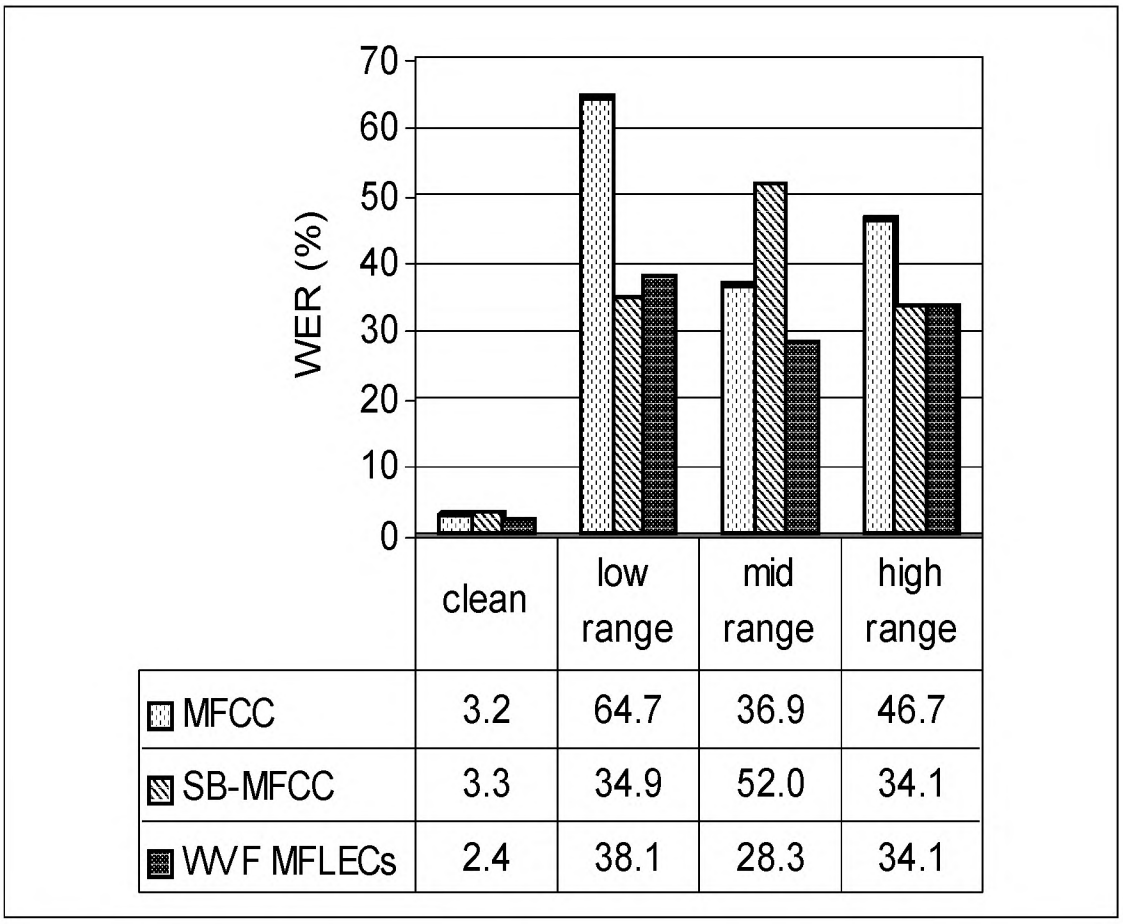

Figure 2: ***WER results for recognition based on the conventional local distance function.***

The WERs obtained for the SB-MFCC based system are summarised in row 2 of Figure 2. The recognition performance for the mid range noise condition shows a much more substantial degradation (compared to the clean condition) than the results obtained for the low and high range distortions,

i.e. the WER increases from about 34% for low and high range noise to almost 52% for mid-range noise. These results indicate that SB-MFCCs loose their so-called *partially smearing* property in the mid-range noise condition. The reason for this loss lies in the fact that the mid-range noise lies at the centre of the mel frequency scale. As a consequence, both sub-bands of the SB-MFCCs are affected by the additive noise. If feature values in both sub-bands are distorted, the spectrally local disturbances within each sub-band are smeared over the whole sub-band. It may therefore be argued that the SB-MFCCs can only prevent the smearing of spectrally local distortions if the distortions are limited to one of the constituent sub-bands. This argument could also be extended to systems with more than two sub-bands, i.e. each sub-band that is affected by noise suffers from the spectral smearing which is inherent in the application of the DCT.

The recognition performance for the clean and the disturbed conditions was also evaluated using a *local distance function with acoustic backing-off*. Figure 3 gives an overview of the results. A comparison of the results in Figure 3 with those in Figure 2 reveals that recognition performance in the disturbed condition is improved for all three feature representations at the cost of some loss in recognition performance in the clean condition. The value of the acoustic backing-off parameter (i.e. the parameter that controls to what extent the contribution to the local distance function is limited for extreme feature values) was chosen based on earlier experience, such that recognition performance in the clean condition did not drop by more than 1.1% absolute (de Veth et al., 1999).

The best overall results are obtained when acoustic backing-off is applied in combination with the WVF-MFLECs feature representation. The WERs for the three different noise conditions showed very little variation, i.e. between 19% and 21%. The consistency in the recognition performance of the WVF-MFLECs system clearly shows that it is less sensitive to the spectral properties of noise than the systems based on SB-MFCCs and MFCCs. Using acoustic backing-off in combination with SB-MFCCs results in significantly better results for both low and high range noise conditions, the WERs are reduced to 32.1% and 29.4% respectively. Even though acoustic backing-off in combination with MFCCs yield significantly better results in low, mid and high range noise conditions, the overall system performance is still worse than the performance of the WVF-MFLEC and SB-MFCC

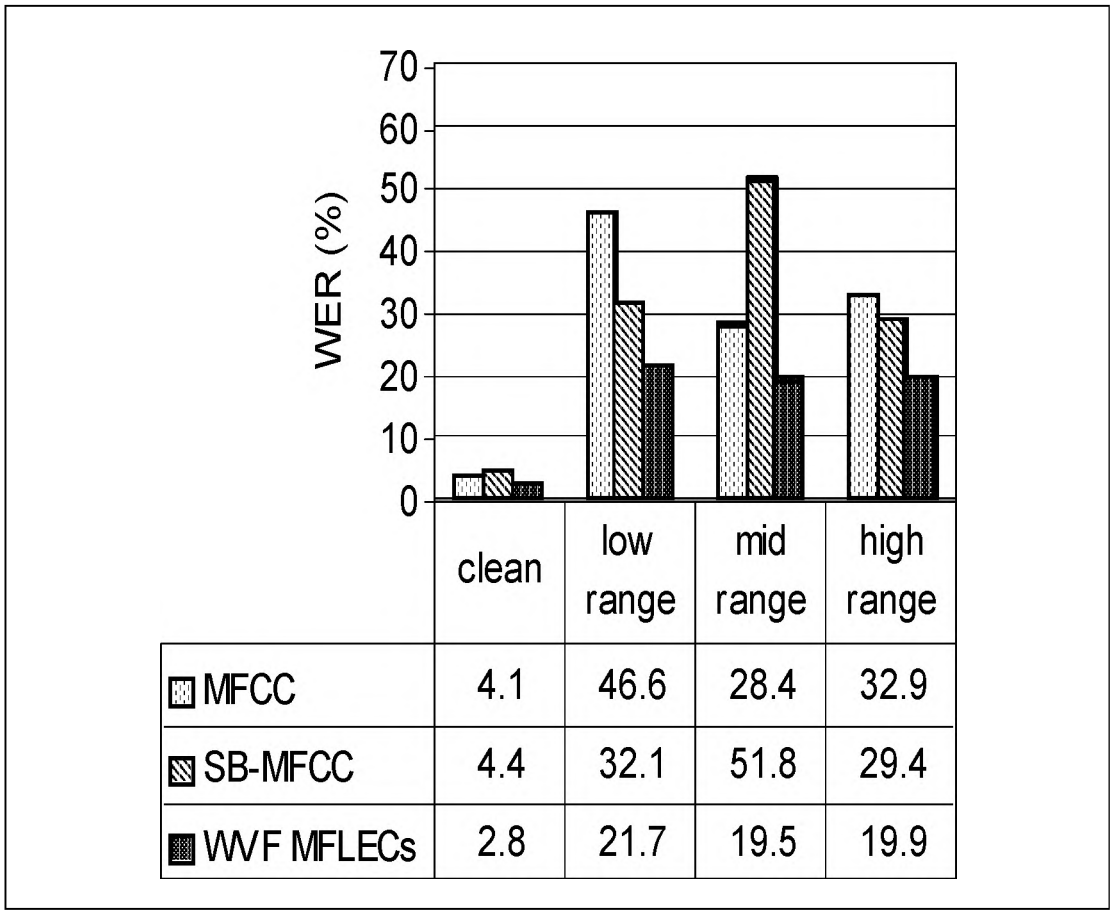

Figure 3: ***WER results for recognition based on a local distance function with acoustic backing-off.***

systems.

The only exception to this statement is in the mid range noise condition where the SB-MFCCs perform much worse than MFCCs. According to the argument presented earlier, SB-MFCCs loose their so-called *partially smearing* property if the effect of the added noise is present in both sub-bands, even to the extent that the application of acoustic backing-off has almost no effect on recognition performance. This observation may seem to suggest that acoustic backing-off is not effective when used in combination with acoustic pre-processing techniques that smear spectrally local distortions over all the components of a feature vector. However, this suggestion is proven wrong by the results in row 1 of Figure 3. Compared to the first

row in Figure 2, these values clearly show that the application of acoustic backing-off in combination with MFCCs leads to a substantial decrease in WER even though MFCCs are known to smear spectrally local distortions over the entire feature vector. Thus the question is: Why does acoustic backing-off seem to work for certain *full smearing* feature representations while it has hardly any effect for others?

A similar effect was reported in a previous investigation (de Veth et al., 1999a). In that study it was shown that the characteristics of the band limited noise used in the experiments affect the individual components of a feature vector differently. For example, the low frequency band-limited noise resulted in an almost equal distortion of all the MFCC components except c_3 and c_9 which were much more severely affected. This uneven distribution may very well explain why the WER is reduced when switching from a conventional local distance function to a local distance function with acoustic backing-off, even on a *fully smeared* set of features like MFCCs. Due to the "outlier" behaviour of coefficients c_3 and c_9, they are likely to be ignored during recognition and will have a diminished impact on recognition performance. The remaining features are all disturbed to some extent which explains why the WER is not fully restored to the level observed in the clean condition.

This observation illustrates that an interaction exists between the characteristics of the noise source and the way these affect the statistical properties of the acoustic features. What needs to be clarified is exactly how noise characteristics are projected into feature space and what the consequences of such a projection are for the feature value distributions. It also illustrates that WER is not always the most suitable way to measure and compare system performance. In order to predict the effectivity of any method used to increase robustness in ASR, knowledge of the distribution of the individual feature vector components that are actually used by the ASR-system, should be taken into account.

Conclusions

In the current experiments, one representation that smears spectrally local distortions over all feature vector components and two representations that limit smearing to a sub-set of the feature vector components were used for modelling and recognition. As a model of distortion, band limited noise was added to speech utterances such that the resulting SNR was 10 dBA. The full smearing representation (MFCC) yielded higher WERs than the representations that keep distortions limited to a subset of vector components (SB-MFCC and WVF-MFLEC). This is a clear indication that, especially in adverse acoustic conditions, care must be taken to choose a feature representation in which possible noise sources affect as few feature vector components as possible.

For most representations investigated in this study, acoustic backing-off appeared to be effective in improving noise robustness. However, the effect of acoustic backing-off on recognition performance is not always what one would expect. For example, the application of acoustic backing-off had hardly any effect in the mid-range noise condition for the SB-MFCC based system while it yielded significantly better results in all noise conditions for the MFCC system even though both systems are based on features that are known to suffer from the effect of so-called *full smearing*. This observation may be explained by the particular way in which the distortions are distributed over the different feature vector components. Some components are much more heavily distorted than others. Acoustic backing-off will limit the impact of the most severely affected outliers (from a statistical point of view), so that recognition is effectively based on those features that are the least affected.

WVF-MFLECs in combination with acoustic backing-off consistently give the best results for all the noise conditions that were studied. WVF-MFLECs are also shown to be less sensitive to the spectral characteristics of noise than the DCT-based representations that were investigated.

Finally, it may be concluded that new methods need to be developed to assess the impact of mismatched training-test conditions on recognition performance. Due to the specific manner a given noise source may affect a

certain type of feature vector, the insights gained from such mismatch assessment tools, are of key importance to further development of noise robust ASR.

Acknowledgement

Part of this research was carried out within the framework of the Priority Programme Language and Speech Technology (TST) which is sponsored by NWO (Dutch Organisation for Scientific Research).

References

Cooke, M., A. Morris and P. Green (1996). Recognising occluded speech. In *Workshop on the Auditory Basis of Speech Perception*, pages 297–300, Keele University, UK. ESCA.

de Veth, J. and L. Boves (1998). Channel normalization techniques for automatic speech recognition over the telephone. *Speech Communication*, 25:149–164.

de Veth, J., B. Cranen and L. Boves (1998). Acoustic backing off in the local distance computation for robust automatic speech recognition. In *Proceedings of ICSLP '98*, pages 1427–1430, Sydney, Australia. ICSLP.

de Veth, J., B. Cranen and L. Boves (1999). Acoustic backing off as an implementation of missing feature theory. http://lands.let.kun.nl/literature/deveth.1999.2.html.

de Veth, J., B. Cranen, F. de Wet and L. Boves (1999a). Acoustic pre-processing for optimal effectivity of Missing Feature Theory. In *Proceedings of Eurospeech '99*, pages 65–68, Budapest, Hungary. ESCA.

de Veth, J., F. de Wet, B. Cranen and L. Boves (1999b). Missing Feature Theory in ASR: Make sure you miss the right type of features. In *Robust Methods for Speech Recognition in Adverse Conditions*, pages 231–234, Tampere, Finland. Nokia, COST 249 & IEEE.

den Os, E. A., T. I. Boogaart, L. Boves and E. Klabbers (1995). The dutch polyphone corpus. In *Proceedings of Eurospeech '95*, pages 825–828, Madrid, Spain.

Dupont, S., H. Bourlard and C. Ris (1997). Robust speech recognition based on multi-stream features. In *Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 95–98, Pont-à-Mousson, France. ESCA & NATO.

Hassall, J.R. and K. Zaveri (1979). *Acoustic noise measurements*. Brüel & Kjær, Denmark.

Lippmann, R. and B. Carlson (1997). Using missing feature theory to actively select features for robust speech recognition with interruptions. In *Proceedings of Eurospeech '97*, pages 37–40, Rhodes, Greece. ESCA.

Morris, A., M. Cooke and P. Green (1998). Some solutions to the missing feature problem in data classification, with applications to noise robust ASR. In *Proceedings of ICASSP '98*, pages 737–740, Seattle, Washington, USA. IEEE.

Nadeu, C., J. Hernando and M. Gorricho (1995). On the decorrelation of filter-bank energies in speech recognition. In *Proceedings of Eurospeech '95*, pages 1381–1384, Madrid, Spain. ESCA.

Noisex (1990). NOISE-ROM-0. NATO: AC243/(Panel 3)/RSG-10, ESPRIT: Project 2589-SAM.

Okawa, S., E. Bocchieri and A. Potamianos (1998). Multi-band speech recognition in noisy environments. In *Proceedings of ICASSP '98*, pages 641–644, Seattle, Washington, USA. IEEE.

Tibrewala, S. and H. Hermansky (1997). Sub-band based recognition of noisy speech. In *Proceedings of ICASSP '97*, pages 1255–1258, Munich, Germany. IEEE.

Young, S., J. Jansen, J. Odell, D. Ollason and P. Woodland (1995). *The HTK Book (for HTK Version 2.1)*. Cambridge University, Cambridge, UK.