ELSEVIER

Speech Communication 34 (2001) 57–74

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Acoustic features and a distance measure that reduce the impact of training–test mismatch in ASR

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

A²RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD, Nijmegen, The Netherlands

Abstract

For improved recognition robustness in mismatched training–test conditions, the application of key ideas from missing feature theory and robust statistical pattern recognition in the framework of an otherwise conventional automatic speech recognition (ASR) system were investigated. To this end, both the type of features used to represent the speech signals and the algorithm used to compute the distance measure between an observed feature vector and a previously trained parametric model were studied. Two different types of feature representations were used: a type in which spectrally local distortions are smeared over the entire feature vector and a type in which distortions are only smeared over part of the feature vector. In addition, two different distance measures were investigated, viz., a conventional distance measure and a robust local distance function in the form of acoustic backing-off. The effects on recognition performance were studied for artificially created, band-limited noise and NOISEX noise added to the speech signals. The results for artificial band-limited noise indicate that a partially smearing feature transform is to be preferred over a fully smearing transform. In addition, for artificial, band-limited noise, a robust local distance function is to be preferred over the conventional distance measure as long as the distorted feature values are outliers with respect to the feature distribution observed during training. The experiments with NOISEX noise show that the combination of feature type and distance measure that is optimal for artificial, band-limited noise is also capable of improving recognition robustness for NOISEX noise, provided that it is band-limited.

Keywords: Automatic speech recognition; Noise robustness; Acoustic features; Missing feature theory; Robust statistical pattern recognition; Robust local distance function

1. Introduction

The present generation of automatic speech recognition (ASR) systems appears to lack robustness when used by real customers to perform real tasks. To a large extent, this lack of robustness is caused by a mismatch between the conditions at recognition time and those in which the ASR system was trained. The 'unexpected' conditions at recognition time may be introduced by user behavior (e.g., coughs, hesitations, repairs, etc.), by the environment (e.g., when the user sits in a room where a radio or TV set is playing, or calls from a noisy train station) or by the transduction and transmission of the signal (e.g., when a cheap microphone integrated in a laptop computer is used or when the signal is transmitted across a fading

* Corresponding author. Tel.: +31-24-3612900; fax: +31-24-3612907.

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

0167-6393/01/$ - see front matter
PII: S0167-6393(00)00046-7

radio path). Additional factors that deteriorate ASR can probably be found by reading the other papers in this issue of Speech Communication.

Regardless of what causes the mismatched training–test conditions, ASR performance deteriorates if there is a mismatch. In addition, ASR performance degrades more as the mismatch becomes more severe. Of course, human speech recognition also suffers in adverse conditions, but it deteriorates at a much slower rate (Lippmann, 1997). Although state-of-the-art ASR algorithms do not intend to mimic human speech processing, it is still worthwhile to try and derive clues from the superior human performance to improve ASR. For some tasks, human superiority is undoubtedly due to the contribution of intelligence and additional sources of information that help decoding. However, humans also appear to outperform ASR systems in tasks where intelligence and linguistic knowledge are of little help, like in the recognition of credit card or telephone numbers (Lippmann, 1997). Therefore, there is ample room for improvement of automatic acoustic decoding under adverse conditions (e.g., Ney, 1999).

In the past two decades, a large number of techniques have been proposed to improve the robustness of ASR in adverse conditions, among which spectral subtraction (Boll, 1979), predictive model compensation techniques (Gales, 1998) and model adaptation techniques (Lee, 1998; Lee and Huo, 1999) are the most popular. In the present paper, we investigate two closely related issues that should enable the improvement of probabilistic modeling and decoding, viz. the type of features used to represent the speech signals and the algorithm used to compute the distance between an observed feature vector and a previously trained parametric model. The research presented in this paper has close links to missing feature theory (MFT) (Cooke et al., 1996; Lippmann and Carlson, 1997; Morris et al., 1998) and robust statistical pattern recognition (Kharin, 1996). Both approaches have a high potential at face value. However, in the context of a conventional ASR system, neither is straightforward to harness so that all their benefits can be exploited. The aim of this paper is to discuss our recent attempts to incorporate some of the principles of MFT and robust statistical pattern recognition in an otherwise conventional ASR system.

2. Principles and hypotheses

In order to be able to explain the hypotheses that we investigated, we first need to discuss two pre-processing steps which are common in conventional ASR systems, i.e., normalization and orthogonalization. Next, we look at the principle that lies at the heart of MFT: distorted observations should not be trusted in the same manner as clean observations (Cooke et al., 1996; Morris et al., 1998). This principle has two immediate consequences: (1) care must be taken to keep clean observations clean, and (2) distorted observations should be treated different from clean observations during the pattern match. The first point leads to a distinction between two different classes of features, viz. a type in which spectrally local distortions are smeared over the entire feature vector and one in which spectrally local distortions are only smeared over part of the feature vector. The second point implies that the conventional way to compute the distance measure should be modified. We will indicate how one of the principles of robust statistical pattern recognition can be used to define a robust distance measure that treats distorted observations different from clean observations. Finally, we combine these more general considerations and formulate the hypotheses that we wanted to investigate in this paper.

2.1. Normalization and orthogonalization

ASR systems represent speech in the form of short-time power spectra and, in many cases, these spectra are obtained through some kind of filter bank operation. In this paper, we will refer to such initial representations of short-time power spectra (i.e., before application of any transform) as the *raw features*. In a conventional ASR set-up, it is considered good practice to train acoustic models from clean speech. If the feature distortions are 'well-behaved' (e.g., time-invariant or known in advance), training can be done using disturbed speech utterances. In all other cases, clean training

data are used and the mismatch between clean training and noisy test conditions must be accounted for explicitly. Even if there is no mismatch between training and test conditions, it is considered good practice to reduce the amount of variation in the data that does not carry important speech information as much as possible. For instance, differences in loudness between recordings are irrelevant for recognition. For reduction of such irrelevant sources of variation, *normalization* transforms are applied. All state-of-the-art ASR systems apply at least some kind of channel normalization (CN), like (cepstral) mean subtraction or (phase corrected) RASTA (de Veth and Boves, 1998) on the raw features. Gain normalization of raw features is also routinely employed (Dautrich et al., 1983).

In addition to feature normalization, most state-of-the-art ASR systems employ some form of feature *orthogonalization*. Orthogonalization is applied because estimates of hidden Markov model (HMM) parameters improve when the ratio of the amount of independent data in the training material to the number of parameters increases. For this reason, many ASR systems assume that the elements in the feature vectors are essentially uncorrelated, so that the covariance matrix becomes diagonal and less parameters need to be estimated. Unfortunately, the raw features (i.e., the log-energy values) are known to be highly correlated. Therefore, it is common practice to apply some kind of orthogonalizing transformation, like the discrete cosine transform (DCT) which yields the well-known cepstral coefficients; other popular orthogonalizing transformations are principal component analysis (PCA) and linear discriminant analysis (LDA) (Hunt et al., 1991).

Conventionally, normalizing and orthogonalizing transformations are combined (cf. mean subtraction after DCT for cepstra). Many of these techniques are capable of improving recognition performance in matched training test conditions. However, none of these techniques, nor any known combination thereof, suffices to maintain recognition performance at the level of the matched condition when a mismatch between training and test conditions occurs that is due to adverse acoustic conditions.

2.2. MFT: keeping clean observations clean

One of the predictions of MFT is that recognition will become more difficult as more features are damaged (Cooke et al., 1996; Morris et al., 1998). Therefore, it should pay to take all possible precautions to avoid unnecessary damage to the features and, if damage cannot be avoided, choose a feature type that concentrates the damage in as few feature components as possible. In this context, it is perhaps surprising that the most popular feature representations used in ASR violate this intuitive maxim. DCT and LDA form linear combinations of all raw features within a vector, thereby smearing distortions in a subset of the raw features over the full vector of transformed features. Moreover, popular forms of pre-processing like CN and computation of delta-features (and delta-deltas) cause distortions which were originally local in time to be smeared over a much wider time window.

In this light, we want to investigate the effects of applying normalization and orthogonalization transformations under adverse conditions. In the present paper, we investigate only the effects of smearing distortions along the frequency axis. We do so by examining four transforms of raw spectral features, viz.

1. within-vector mean normalized Mel-frequency log-energy coefficients (indicated as F1);
2. Mel-frequency cepstral coefficients (indicated as F2);
3. sub-band Mel-frequency cepstral coefficients (indicated as P1) (Okawa et al., 1998);
4. within-vector filtered Mel-frequency log-energy coefficients (indicated as P2) (Nadeu et al., 1995).

Details of these feature representations will be given in Section 3. For the moment, it suffices to note that the first two of these representations (F1 and F2) are calculated from the entire vector of raw features. Therefore, a distortion present in any one raw feature will affect all transformed features. The label 'F' for this type of feature reflects the fact that the transform *fully* smears distortions. The last two representations (P1 and P2) are designed to limit the extent to which a distortion in some part of the raw feature vector spreads over

the entire transformed feature vector. The label 'P' for this type of features reflects the fact that the transform *partially* smears distortions.

It is reasonable to expect that the *number* of distorted feature values is not the only factor determining the degradation of recognition performance under adverse conditions. The *degree* of distortion is probably equally important. However, it is not possible to predict the degree to which the transformed features will be affected without prior knowledge of the details of the distortion of the raw features (i.e., which components are affected and exactly how large the distortion is). In this paper, we discuss a number of experiments that allow us to illustrate the interaction between the type of transformation used for computing the final acoustic features and the nature of the distortion. To this end, we investigate the effect of artificially created, band-limited noise when added at three different signal-to-noise ratios (SNRs) and in three different frequency regions. In addition, we study the effect of adding three different noise types taken from NOISEX (Noisex, 1990). More details about the distortions together with motivations for our choices are given in Section 3.

2.3. Robust distance computation as an implementation of MFT

According to MFT, feature values that are highly corrupted are best discarded when computing the distance between an observation vector and a set of models (Cooke et al., 1996; Lippmann and Carlson, 1997; Morris et al., 1998). Although the idea that garbage should not be treated as data may be easily understood, telling the garbage and the data apart in noise-robust ASR is not an easy task. Over the last years, considerable effort has been spent on making the distinction between disturbed and undisturbed raw features by operations on the log-energy values, without using prior knowledge about the particular speech sound under consideration (Dupont et al., 1997; Tibrewala and Hermansky, 1999; Vizinho et al., 1999). Recently, good results were reported in an experimental set-up where artificial noise was added to wide-band speech (Vizinho et al., 1999). However, it is still an open question how these results will carry over to a task in which the available frequency band is much smaller (e.g., telephone or GSM speech). In addition, few explicit speech–noise distinction methods exploit the fact that the sounds to be told apart from noise are speech sounds. It is well known that the spectra of individual speech sounds show a high degree of variability. Therefore, to distinguish between disturbed and undisturbed feature values, the variability of the spectrum of the speech sound under analysis should ideally be taken into account.

In (de Veth et al., 1998, 1999, 2001), a new method was proposed in which hard decisions about which feature vector components are disturbed or undisturbed are avoided. This new method is based on a technique of robust statistical pattern recognition (Huber, 1981; Kharin, 1996). According to one of the principles of robust statistical pattern recognition, the tails of any distribution of observed values are inherently difficult to estimate reliably because tails contain few data points. A speech utterance produced under adverse acoustic conditions may contain a proportion of feature values that are outliers with respect to the distributions derived from the training speech. One way to limit the effect of these noise-induced outliers is to model each set of observations by means of two distributions: one that is obtained from the training data and another that represents all feature values not observed in the training material, i.e., the outliers. With this approach, the likelihood of observing a feature vector in a state can be described as an interpolation between the likelihood yielded by the trained distribution and the likelihood derived from the distribution that represents the values not previously seen. The interpolation weights should reflect the expected severity of the distortions.

In (de Veth et al., 1998, 1999, 2001) the distribution of the 'non-observed' values was assumed to be uniform, and this approach was called *acoustic backing-off*. By choosing the uniform distribution, it is ensured that the shape of the combined distribution will still have much in common with the shape of the 'clean' distribution observed during training. It was shown (de Veth et al., 2001) that acoustic backing-off can be

considered as an implementation of MFT that (1) is suitable for use in a conventional ASR system, (2) is suitable for use with any feature representation and (3) wires the detection mechanism for identifying disturbed feature vector elements into the decoder. In this paper, we will compare the conventional local distance measure with our robust local distance computation in the form of acoustic backing-off.

2.4. Hypotheses under investigation

In short, this paper investigates four different feature transformations (two F-type and two P-type features), two different methods for local distance computation (conventional versus robust implemented in the form of acoustic backing-off) and several different noise types (band-limited noise added at several SNRs and, more importantly, at different frequency locations and three more realistic noise types) in their mutual interaction. By studying the effects of different combinations of feature type, local distance and noise type, we wanted to test the following hypotheses:

1. A transform that smears local distortions over the complete feature vector (an F-type transform) is more vulnerable to spectrally local distortions than a transform that keeps local distortions local (a P-type transform).
2. Hypothesis 1 holds true irrespective of the distance measure used: conventional or robust.
3. A robust local distance measure outperforms the conventional distance measure.
4. The performance gained by a robust local distance measure is larger if the distortions are limited to a smaller number of features.
5. Hypotheses 1–4 hold irrespective of the type of distortion, as long as the spectrum of the noise is confined to a limited frequency range.

These hypotheses focus on the conditions for which improvement of recognition robustness may be expected. To evaluate our hypotheses as functions of the type of noise, three experiments were performed. In the first two experiments, corrupted speech was obtained by adding artificially created, band-limited noise to clean speech signals. Artificial noise was used in these experiments because it facilitated testing our hypotheses. In the third experiment, we used car, babble and factory noise taken from the NOISEX database (Noisex, 1990) to create corrupted speech signals. This experiment should provide an impression of the capabilities of acoustic backing-off for noise types that are more realistic and not necessarily band-limited.

The rest of this paper is organized as follows. First, in Section 3, we describe the experimental set-up that we used in more detail. In Section 4, we compare the recognition performance for the four different types of features. We evaluated system performance with clean and disturbed data for each of the four acoustic representation techniques, with and without applying MFT in the form of acoustic backing-off. In addition, we present results for the three different noise types taken from NOISEX. In Section 5, we propose interpretations for the results of our experiments. Finally, our conclusions are presented in Section 6.

3. Experimental set-up

3.1. Speech material

The speech material for our 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. Speech signals were recorded from a primary rate ISDN telephone connection. 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. For training we used a set of 1997 strings (16,582 digits). Care was taken to balance the training material with respect to (1) an equal number of male and female speakers, (2) an equal number of speakers from each of the 12 provinces in the Netherlands and (3) an equal number of tokens per digit. For cross-validation during training (cf. de Veth and Boves, 1998) we used 504 digit strings (4300 digits). All the models were evaluated with an independent set of 1008 test utterances (8300 digits). The cross-validation test set and the independent test set were balanced according to the same criteria as the training material. None of the original utterances used for

training or testing had a high background noise level.

3.2. Acoustic features

Each of the four different acoustic feature representations in our experiments was based on a transformation of the raw feature representation that consisted of 16 Mel-frequency log-energy coefficients (MFLECs). The MFLECs were computed using a 25 ms Hamming window shifted with 10 ms steps and a pre-emphasis factor of 0.98. Based on a Fast Fourier Transform, 16 filter band energy values were calculated, with the filter bands 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). In addition to the 16 MFLECs, we also computed the total log-energy for each frame. These signal processing steps were performed using HTK2.1 (Young et al., 1995).

For the F1 features, we computed the average within-vector log-energy value for each frame. This within-vector average was subtracted from each of the original 16 MFLEC values yielding 16 within-vector mean normalized MFLECs. CN was accomplished by subtracting the average value (computed over the whole utterance) for all 16 F1 values. Finally, we computed the (smoothed) time derivatives (delta-coefficients). Combining these with the 16 static values, log-energy and delta log-energy we obtained 34-dimensional feature vectors.

In the case of F2 features, Mel-frequency cepstra $\{c_1, \ldots, c_{12}\}$ were computed from the raw MFLECs using the DCT. CN was done by means of cepstrum mean subtraction (CMS) over the entire utterance. Finally, we computed the time derivatives and added these to the 12 channel-normalized Mel-frequency cepstral coefficients. Together with log-energy and delta log-energy, we obtained 26-dimensional transformed feature vectors.

The P1 values were obtained by computing $\{c_{1,1}, \ldots, c_{1,6}\}$ independently for the first eight MFLEC values (covering 0–1218 Hz) and $\{c_{2,1}, \ldots, c_{2,6}\}$ for the remaining eight MFLECs (covering 1015–4000 Hz). We used two sub-bands with eight MFLEC values each for computing the P1 features because this was reported to be the optimal feature recombination in (Okawa et al., 1998). Next, we proceeded exactly as with the Mel-frequency cepstral coefficients, i.e., subtracting the mean computed over the whole utterance for CN and computing the deltas. Together with log-energy and delta log-energy, we arrived at 26-dimensional feature vectors.

The P2 values were computed by applying the filter $z - z^{-1}$ within each frame for coefficients 2–15. In other words, the spectral difference $x(\omega_{k+1}, t) - x(\omega_{k-1}, t)$ was computed for $k = 2, \ldots, 15$, where $x(\omega_k, t)$ denotes the MFLEC value computed for filter band k at time t. MFLECs 1 and 16 were simply copied. After this combination of differencing and copying, the mean value of each vector component was computed over the whole utterance and subtracted as a form of CN. Next, the deltas were computed. The static and delta within-vector filtered MFLECs were combined with log-energy and delta log-energy to arrive at 34-dimensional feature vectors.

3.3. Additive noise

In order to test our hypotheses discussed in Section 2, we started with additive band-limited Gaussian noise as a distortion. In the first experiment, low frequency noise was added at decreasing SNRs of 20, 10 and 5 dBA (both the speech and noise energy levels were weighted according to the A-scale (Hassall and Zaveri, 1979)). The different SNRs allow us to study the impact of the severity of the distortion; at the same time they allow us to investigate whether there is a qualitative difference between full-smearing and partial-smearing transformations. Moreover, comparing the performance of a conventional and a robust local distance function should shed light on the conditions for which the robust approach is superior.

The band-limited noise signals were obtained by filtering Gaussian white noise signals using a fifth-order elliptical filter. The cut-off frequencies of the band-pass filter were chosen so that approximately one quarter of the raw features would be contaminated: $F_{\text{low}} = 395$ and $F_{\text{high}} = 880$ Hz.

The high cut-off frequency of the noise was chosen so that noise distortions were present only in the first set of energy bands used in the P1 feature representation. Therefore, this set-up allows us to investigate whether a qualitative difference exists between the two full-smearing feature representations on the one hand and the two partial-smearing feature representations on the other.

Fig. 1 illustrates how this type of distortion affects the different feature representations used in this study. The values of the static features (i.e., excluding log-energy, the delta-coefficients and delta log-energy) are shown as functions of time before (first column) and after (second column) adding noise at 10 dBA SNR to an utterance /een vier/ (i.e., one four). In the third column, the normalized squared error NSE_{kl} is shown, which is defined as

$$\mathrm{NSE}_{kl} = \frac{(x'_{kl} - x_{kl})^2}{\sum_{l=1}^{L} x_{kl}^2}, \tag{1}$$

where x'_{kl} denotes the disturbed value of coefficient k at time l, x_{kl} the value of the undisturbed coefficient and L the total length of this utterance. As a reference, the top row of Fig. 1 shows original, disturbed and NSE values for the 16 raw

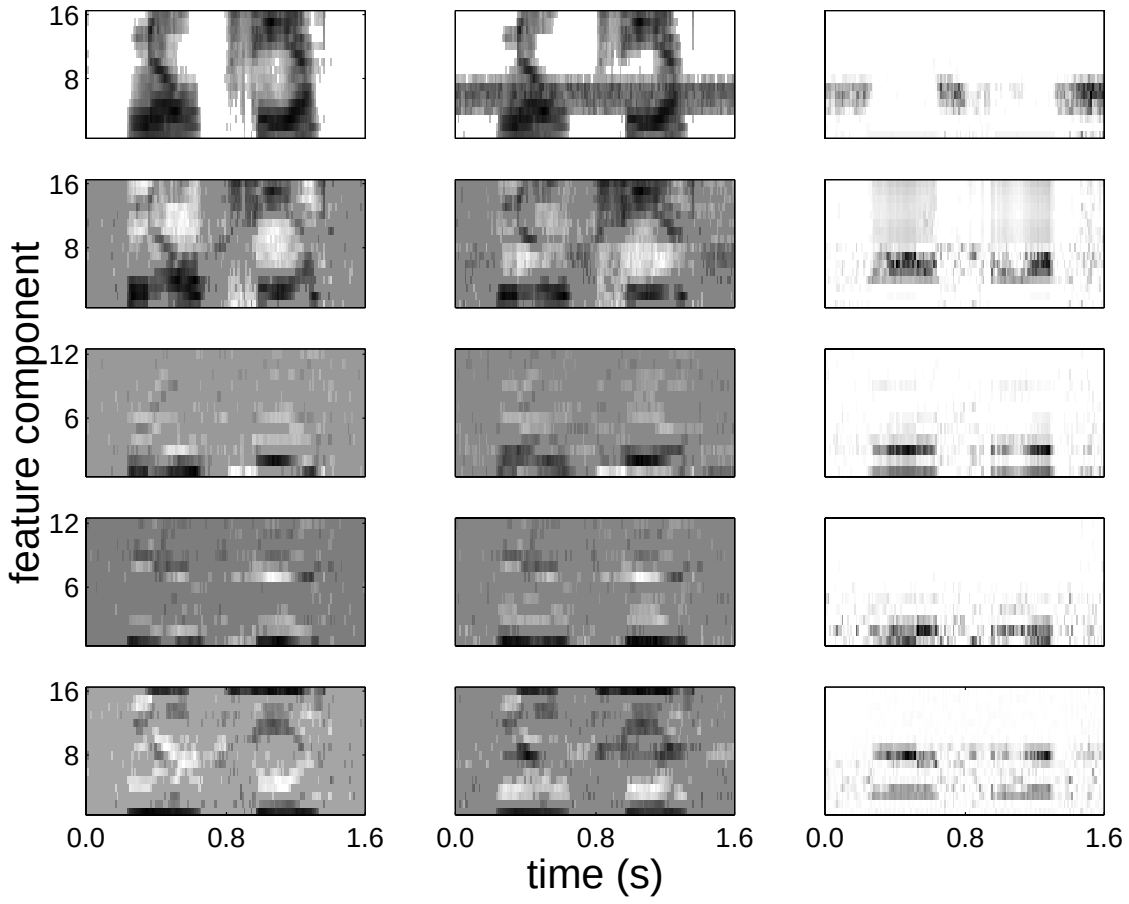

Fig. 1. First column: Parameter tracks of the static features for a clean utterance /een vier/ (i.e., one four). Second column: Parameters for the same utterance after addition of band-limited noise with SNR = 10 dBA. Third column: Normalized squared error as defined in Eq. (1). The first row corresponds to the 16 raw Mel-frequency log-energy coefficients. Rows 2–5 correspond to feature types F1, F2, P1 and P2, respectively.

MFLECs. Rows 2–5 correspond to feature types F1, F2, P1 and P2, respectively. To enhance visual clarity of the speech regions, we applied a non-linear mapping to the gray-scale for each plot in the first and second columns, so that the gray level in non-speech regions becomes almost uniform. However, for each error plot in the third column, a straightforward linear mapping of the gray scale was used, since visibility enhancement of particular parts in the error plots was deemed unnecessary in these cases.

Looking at Fig. 1, the following observations can be made. Firstly, the two spectrograms in the top row (first and second columns) clearly show the regions where speech is present. The spectrogram for the distorted condition clearly shows that the additive noise is restricted to a limited number of bands, just as we intended. The error plot of the raw features (top row, third column) shows that the main differences are located in the portions of the signal that were silent in the clean condition (i.e, at the beginning and end of the utterance and in between the two digits). This is exactly as expected from a log-energy measure. This panel also shows that the energy of the band-limited noise is masked during most of the speech portions of the signal. Secondly, for the four feature representations shown in rows 2–5, the original and distorted feature plots are more difficult to interpret. Differences between the original and distorted feature plots can be discerned. However, the differences are not as easily spotted as in the spectrograms shown in the top row. As a result, the error plots appear to be a more appropriate source of information about the distortions in these cases. It can be seen in the error plots shown in rows 2 and 3 that the distortions in the F1 and F2 features are smeared out over a larger proportion of the coefficients compared to the raw features shown in the top row. In addition, there are only very few distortions visible in the top half of the coefficients for the P1 and P2 features (rows 4 and 5) for this type of distortion. These findings illustrate that band-limited, artificially created noise leads to qualitatively different error plots for the 'F'- and 'P'-type features. Therefore, this type of noise is highly suitable for testing our hypotheses about 'F'- and 'P'-type features.

In the second experiment, we studied the effects of adding band-limited Gaussian noise to the speech signals in three different frequency regions. For the first frequency region, which we will refer to as the 'low range', we used exactly the same cut-off frequencies as in the first experiment: $F_{\text{low}} = 395$ and $F_{\text{high}} = 880$ Hz. For the second region ('mid-range' noise), we used $F_{\text{low}} = 833$ and $F_{\text{high}} = 1446$ Hz. Finally, for the third region ('high-range' noise), we used $F_{\text{low}} = 1446$ and $F_{\text{high}} = 2303$ Hz. In all cases, the SNR level was kept constant at 10 dBA. These noises allow us to study the interaction between the frequency region of the distortion, the transformation of the raw features and the type of local distance function.

Finally, in the third experiment, we added 8 kHz downsampled versions of NOISEX car, babble and factory noise to our test signals. In all cases, the noise amplitude level was adjusted to obtain an SNR of 10 dBA. To assess the frequency regions that are most affected by these three types of noise, the power spectrum computed over a noise segment of 1 s is shown in Fig. 2 for each noise type.

It can be seen in Fig. 2(A) that car noise is band-limited with energy predominantly in the frequency region below 200 Hz. Figs. 2(B) and (C) indicate that the spectra of babble and factory noise, respectively, carry significant amounts of energy at all frequencies. Thus, these two noise types are not band-limited; these noises are wide-band.

3.4. Hidden Markov modeling

The 10 Dutch digit words were described with 18 context-independent phone models. In addition, we used three different models for silence, background noises and out-of-vocabulary speech. Each phone unit was represented as a left-to-right HMM consisting of three states, with the emission pdf of each state in the form of a single Gaussian pdf and only self-loops and transitions to the next state. For these models, the total number of states was 63 (54 for the phones plus nine for the silence and noise models). We used HTK2.1 for training and testing HMMs (Young et al., 1995). We followed the cross-validation scheme described in

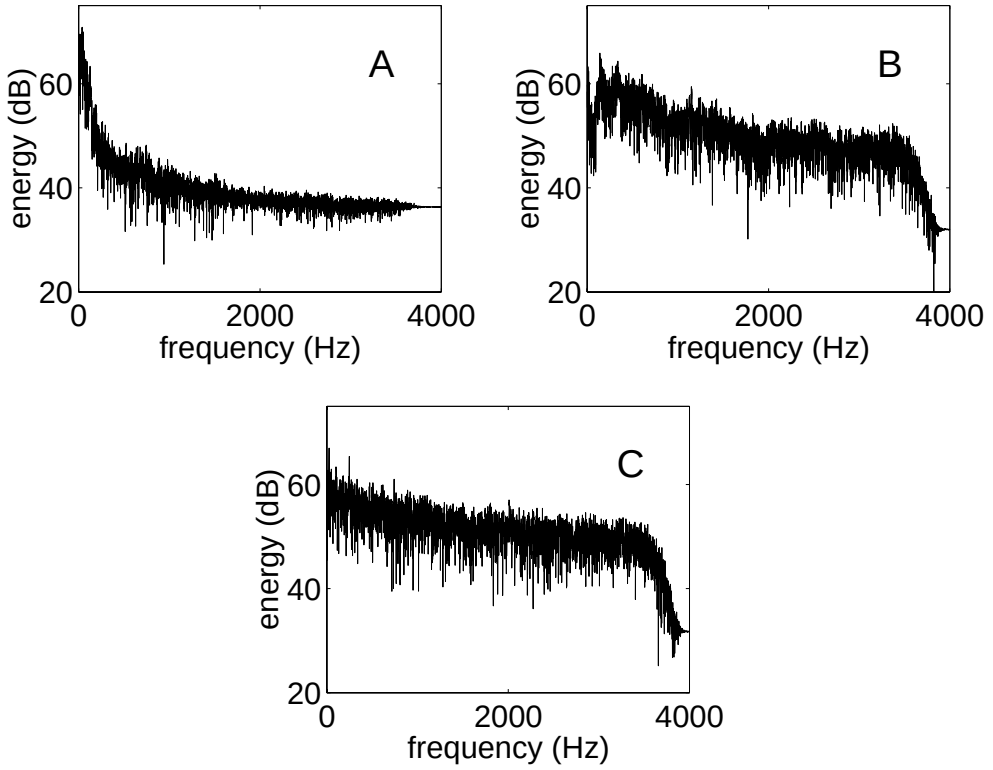

Fig. 2. Long-term spectrum of NOISEX (A) car noise; (B) babble noise; (C) factory noise.

(de Veth and Boves, 1998) to determine the optimal number of Baum–Welch iterations. The eventual models were obtained through subsequent mixture splitting. We split up to four times, resulting in recognition systems with 16 Gaussians per state (containing 1008 Gaussians in total). We used diagonal covariance matrices for all HMMs and each model set was trained only once, using undisturbed features. The recognition syntax used during cross-validation and testing was defined, so that connected digit strings varying in length from 3 to 16 digits could be recognized.

3.5. Robust local distance function

Consider an HMM state S_i that is described by a mixture of M Gaussian probability density functions with diagonal covariance matrices. The conventional local distance function d_{loc} is equal to the sum of the emission cost associated with that state and the transition cost from the previously visited state to state S_i. When the transition cost is disregarded, the local distance function becomes equal to the emission cost associated with state S_i:

$$d_{\text{loc}}(S_i, \boldsymbol{x}(t)) = -\log\left\{\sum_{m=1}^{M} w_{im} \prod_{k=1}^{K} G_{imk}(x_k(t))\right\}, \tag{2}$$

where $\boldsymbol{x}(t)$ denotes the acoustic observation vector at time t, w_{im} the mth mixture weight for state S_i, K the dimension of the acoustic observation vector, $x_k(t)$ the kth component of $\boldsymbol{x}(t)$ and G_{imk} the kth component of the mth Gaussian probability density function for state S_i. The robustness of the statistical distance measure defined in Eq. (2) can be improved, when outlier values occur, by replacing the conventional Gaussian probability density function with a robust probability density function that corresponds to a mixture of two

statistical processes: the process describing the observations that were seen during training and the process describing the values that were not previously seen (Huber, 1981; Kharin, 1996). According to this idea, a robust local distance function d_{robust} was defined in (de Veth et al., 1998, 2001) as

$$d_{\text{robust}}(S_i, \boldsymbol{x}(t)) = -\log\left\{\sum_{m=1}^{M} w_{im} \prod_{k=1}^{K} [(1-\epsilon) G_{imk}(x_k(t)) + \epsilon p_0(x_k(t))]\right\}, \quad (3)$$

where ϵ denotes the a priori probability that a feature value originates from the process of unseen feature values ($0 \leqslant \epsilon < 1$) and $p_0(\cdot)$ the probability density function associated with the unknown process. It can be seen that Eq. (3) immediately reduces to Eq. (2) if we choose $\epsilon = 0$.

In order to be able to actually use Eq. (3), we need to decide how the process for the unseen observations is best described, where 'best' means optimal according to the theory of robust statistical pattern recognition (Kharin, 1996). For the particular problem we study (i.e., how to make the computation of the local cost in the search robust), the best description of the unknown process is, as yet, an open question. In (de Veth et al., 1998, 2001), we proposed to model the unknown process with a uniform distribution. This distribution is simple, reflects our assumption that we do not have any prior knowledge about the process underlying the unseen observations and is capable of improving recognition robustness for artificial distortions. The technique of using a robust local distance function with a uniform distribution for $p_0(\cdot)$ was named acoustic backing-off (de Veth et al., 1998, 2001).

Having decided to describe the process for the unseen observations with a uniform distribution, we must still choose a value for the a priori probability ϵ that a feature value originates from the process of unseen observations. Ideally, the optimal value should depend on the severity of the distortions. We do not know yet how to quantify 'severity' in terms of a measure that is correlated with ϵ. 'Severity' will depend on the type of distortion, so eventually it might become possible to narrow down the range of reasonable values when the most important characteristics of the noise are known. We also do not know how ϵ can be estimated from a speech utterance that must be decoded. However, for the noise types used in (de Veth et al., 1999, 2001), it was found that recognition performance is not critically sensitive to variations of this parameter if $\epsilon \approx 0.1$. In all experiments reported in this paper, we used $\epsilon = 0.1$.

4. Results

4.1. Baseline recognition performance

In order to determine a proper reference system for each feature representation, we trained HMMs according to the cross-validation procedure described in (de Veth and Boves, 1998). For the resulting model sets, we determined the word error rate (WER) using the 1008 utterances of the test set. The WER was defined as

$$\text{WER} = \frac{S + D + I}{N} \times 100\%, \quad (4)$$

where N is the total number of words in the test set, S the total number of substitution errors, D the total number of deletion errors and I the total number of insertion errors. For the recognition experiments, we used HMM systems with 16 Gaussians per state. The WER results obtained at this working point are shown in Table 1 for the four different feature sets that were studied (the figures in brackets indicate the 95% confidence intervals). As can be seen in Table 1, WER values were obtained in the range of 2.4% (P2) to 3.4% (F1). The WER values of F1, F2 and P1 do not

Table 1
WER results for four different feature representations in the clean condition. The figures in brackets indicate the 95% confidence intervals

Feature representation	WER
F1	3.4 (0.4)
F2	3.2 (0.4)
P1	3.3 (0.4)
P2	2.4 (0.3)

show significant differences. However, the P2 representation does yield significantly better results. This finding agrees with observations reported in (Nadeu et al., 1995).

4.2. Experiment 1

Using the distortion of 'low-range' noise at SNR levels of 20, 10 and 5 dBA, we evaluated system performance for a conventional local distance function. In addition, we evaluated the recognition performance for our acoustic backing-off implementation of the robust local distance function. The results for the conventional local distance function are shown in Fig. 3. The WER results using the robust local distance function are shown in Fig. 4(A). Fig. 4(B) shows the WER difference, $\Delta\mathrm{WER} = \mathrm{WER}_{\mathrm{robust}} - \mathrm{WER}_{\mathrm{conventional}}$. Thus, in Fig. 4(B), a negative value indicates a recognition improvement and a positive value a deterioration.

Looking at the WER differences in the clean condition (the leftmost group of bars in Fig. 4(B)), it can be seen that a small loss in recognition performance occurs when switching from the conventional to the robust local distance function.

Figs. 3 and 4 show that recognition performance is better for the two feature representations that only partially smear distortions (P1 and P2; the two rightmost bars) than for the two that

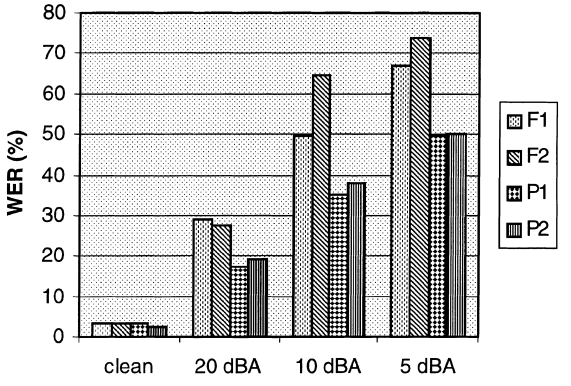

Fig. 3. Recognition results as a function of signal-to-noise ratio when using the conventional local distance function. F1: within-vector mean normalized Mel-frequency log-energy coefficients; F2: Mel-frequency cepstral coefficients; P1: sub-band Mel-frequency cepstral coefficients; P2: within-vector filtered Mel-frequency log-energy coefficients.

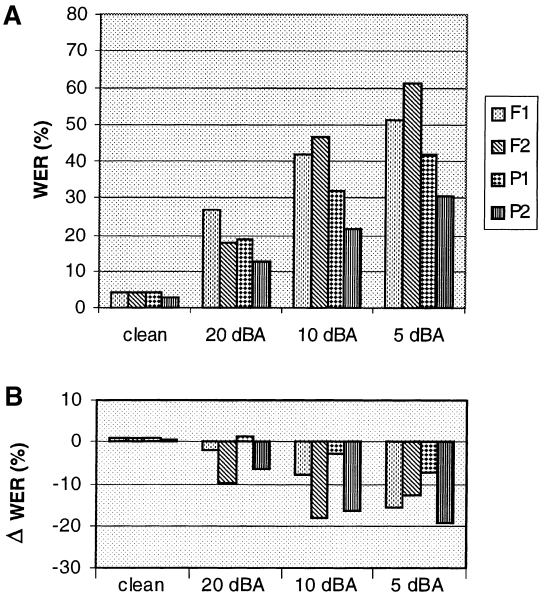

Fig. 4. (A) Recognition results as a function of signal-to-noise ratio when using the robust local distance function. (B) Corresponding ΔWER values.

smear distortions over all feature components (F1 and F2; the two leftmost bars). This observation holds both for the recognizer with the conventional (Fig. 3) and for the one with the robust local distance function (Fig. 4(A)). Second, the recognizer with acoustic backing-off yields better results in the noisy conditions than the recognizer with a conventional local distance function in all but one case. As can be seen in the second set of bars in Fig. 4(B), the single exception occurs for the P1 features at SNR = 20 dBA for which a small (but statistically significant) performance degradation can be observed: The WER increased with 1.5% from 17.1% to 18.6% when switching from the conventional to the robust local distance function. In all other cases with noise present, the robust local distance function improved the performance, although the gain is occasionally small (cf. the results for F1 features at SNR = 20 dBA and for P1 features at SNR = 10 dBA).

4.3. Experiment 2

For each feature type, we evaluated system performance for low-, mid- and high-range noise.

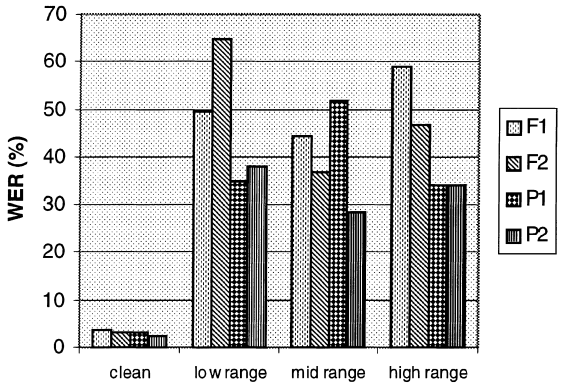

Fig. 5. Recognition results as a function of the frequency region of the noise when using the conventional local distance function.

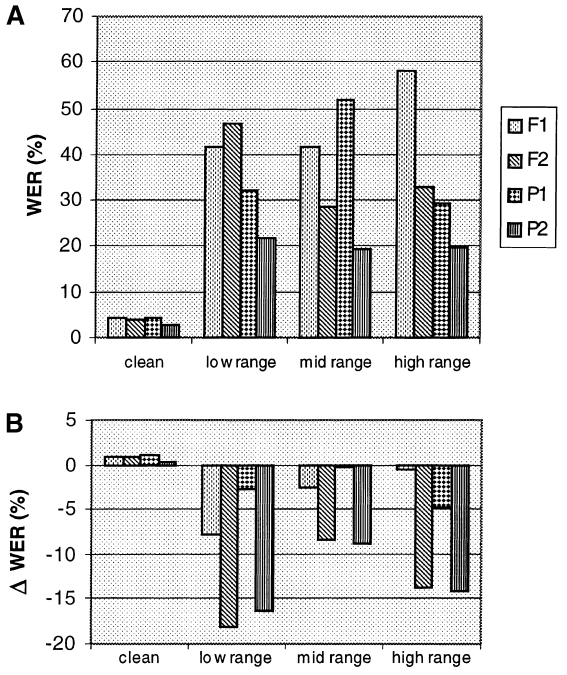

Fig. 6. (A) Recognition results as a function of the frequency region of the noise when using the robust local distance function. (B) Corresponding ΔWER values.

In each noise condition, the SNR was 10 dBA. The results for the conventional local distance function and for the robust local distance function are shown in Figs. 5 and 6, respectively.

The results for the conventional local distance computation show that all feature representations appear to be sensitive to the spectral location of noise, although there are clear differences in the sensitivity patterns for the different feature types. The WER appears to be lowest for the mid-range noise for F1, F2 and P2 features. However, for P1 features, the pattern appears to be the opposite: recognition performance is worst for the mid-range noise. In addition, it can be seen that both 'F'-type features outperform the P1 features for the mid-range noise. For the F2 features, this finding is in good agreement with some of the results reported in (Okawa et al., 1998). In that study, it was also found that Mel-frequency cepstral coefficients outperformed sub-band Mel-frequency cepstral coefficients for the NOISEX-92 noise types 'buccaneer2', 'leopard' and 'white'. For these noise types, as well as for our mid-range noise, energy is present in both sub-bands. With distortions in each sub-band, not a single cepstral coefficient of the P1 features will remain unaffected. As a consequence, the potential advantage of the P1 features over the F2 features is annihilated. This reasoning explains why P1 does not have an advantage over F2. However, it does not explain why the P1 results for mid-range noise are so much worse compared to the F2 results.

When we compare the recognition results obtained with the robust local distance function to the conventional results (see Fig. 6(B)), we observe that WER is reduced in noise conditions, although the reduction was very small in some cases (i.e., for P1 features with mid-range noise and for F1 features with high-range noise). Secondly, it can be seen that the pattern of WER sensitivity to the spectral location of noise does not change for all four feature representations. For the combination of P2 features and robust local distance function, the WERs are almost on the same level for different locations of noise. This means that this combination of feature representation and local distance function seems to be almost insensitive to the location of the noise.

4.4. Experiment 3

In order to get an impression of the capacity of acoustic backing-off to improve recognition performance for more realistic noise conditions, three

Table 2
WER results for four different noise conditions using the P2 features. The figures in brackets indicate the 95% confidence intervals

Noise condition	Conventional LDF	Robust LDF
No noise	2.4 (0.3)	2.8 (0.4)
Car (10 dBA)	9.3 (0.6)	4.8 (0.5)
Babble (10 dBA)	30.4 (1.0)	28.8 (1.0)
Factory (10 dBA)	31.5 (1.0)	32.7 (1.0)

experiments were conducted in which NOISEX car, babble and factory noises (Noisex, 1990) were added to the test signals at a level of 10 dBA. In these experiments, we limited ourselves to the feature representation that yielded the best overall results for artificial, band-limited noise, i.e., the P2 features. The WER results are listed for the conventional and the robust local distance function in Table 2.

The results in Table 2 indicate that the effectiveness of acoustic backing-off depends on the type of noise. For car noise, the improvement is statistically significant and the WER is reduced by 48% when switching from a conventional to a robust local distance function. However, for babble noise, there is hardly any reduction and the WER deteriorates for factory noise. Upon checking the result for factory noise, we found a WER of 31.4% (i.e., essentially the same value as found for the conventional local distance function) when the value for ϵ was actually optimized for this type of noise. Thus, a robust local distance function implemented in the form of acoustic backing-off leaves the recognition performance at best unaffected for factory noise. We recall from Section 3 that the long-term spectra of babble and factory noise contain significant contributions at all frequencies in the frequency range of 0–4 kHz (see Figs. 2(B) and (C)). Apparently, the combination of P2 features and acoustic backing-off is not capable of enhancing recognition robustness for wide-band noise.

5. Discussion

With our experiments, we wanted to test several hypotheses. First, we hypothesized that a transformation of raw features that smears distortions that are only present in a restricted frequency region over the full transformed vector is inherently more susceptible to performance degradation under adverse conditions than a transformation that keeps local distortions local. To that end, we used two 'full-smearing' transformations (F1 and F2) and two 'partial-smearing' ones (P1 and P2) in combination with band-limited noise.

The results of our experiments seem to confirm this hypothesis, although, perhaps, not without qualifications. In experiment 1, with low-range noise added at several SNRs (cf. Fig. 3), the two full-smearing feature sets always show substantially higher error rates. The ranking of the four feature sets differs somewhat between the four SNR conditions. This should warn against rash conclusions about inherent advantages of a specific feature set.

The results of experiment 2, in which noise was added to speech at different spectral locations, also support the hypothesis that partial-smearing transforms are preferable. The fact that the P1 features lose their advantage in the case of mid-range noise may seem a contradiction, but is easy to explain. The P1 features can only be considered as partially smearing if distortions are limited to one of the sub-bands. The low-range and high-range noise conditions were constructed to accomplish this. Indeed, in these conditions, P1 is among the better features. The mid-range condition was constructed to see what happens if both sub-bands are affected. It is clear that P1 features are not robust if distortions are present in both sub-bands. This finding casts doubt on the value of sub-band cepstra as representations of the speech signal that should be robust against arbitrary distortions.

Overall, the partial-smearing P2 features seem to be the best choice, at least for the somewhat artificial noise conditions investigated in experiments 1 and 2. To some extent, their gain relative to the much more popular F2 features may be due to the fact that the P2 models had more parameters (recall that there were 34 P2 features and only 26 F2 features). However, in the context of the theory to which this paper intends to contribute, P2 features may be inherently superior: according

to the error plots in the third column of Fig. 1, P2 features (row 5) show a lesser degree of smearing of band-limited noise compared to F2 features (row 3).

The hypothesis that a robust local distance function is always to be preferred over the conventional one appears to be confirmed by the results of all experiments where band-limited noise was used as a distortion, i.e., irrespective of whether artificially created noise or more realistic noise was tested. For band-limited noise, the only systematic exceptions where a conventional distance measure is preferred over a robust local distance are found in the case of clean data, i.e., the condition where training and test match completely. In our future research, we plan to investigate the option of using robust statistics in model training too, to see if that would remove the disadvantage under matched conditions.

With respect to the hypothesis that the advantage of acoustic backing-off would be greater if less feature components are affected (e.g., with P-type features rather than F-type ones), our results appear to be somewhat less convincing. Actually, the results in this paper seem to contradict the results we found in earlier experiments (de Veth et al., 1999, 2001). In those experiments, we did indeed find that acoustic backing-off had no positive effect in combination with F2 features. This difference in behavior may be explained by the fact that the distortions used in (de Veth et al., 1999, 2001) differ from the noise types used in the present study. Summing up, the experiments with band-limited noise show that (1) acoustic backing-off can give an advantage for all feature sets (including F2 features), (2) the gain is not always equally large, and (3) the gain is sometimes larger than we expected. For wide-band noise, however, the robust local distance function did not outperform the conventional distance measure, at least not for P2 features.

The fact that acoustic backing-off can give an advantage for both types of acoustic features investigated in this paper (i.e., F- and P-type) can be explained by a detailed analysis of the impact of distortions on transformed feature sets of test data. For the purpose of this analysis, the normalized mean squared error (NMSE) of corresponding values of disturbed and undisturbed feature components was used (Huerta and Stern, 1998). The NMSE is computed for each feature component k as

$$\mathrm{NMSE}(k) = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L_n}(x'_{kln} - x_{kln})^2}{\sum_{n=1}^{N}\sum_{l=1}^{L_n}x_{kln}^2}, \tag{5}$$

where N is the total number of utterances used for computation, L_n the length of each individual utterance n, x'_{kln} the value of the disturbed feature component k in observation vector l of utterance n and x_{kln} the original undisturbed feature value. As can be seen in Eq. (5), the NMSE of each individual feature component k is the ratio between the average energy of the distortions and the average energy of the original feature values, where the average is taken over all observed feature vector sequences. In a set-up where both the clean and the disturbed utterances are available, the NMSE can be readily computed.

To illustrate how the band-limited noise affects F2 feature coefficients, the NMSEs are shown in Fig. 7 for the condition with SNR = 10 dBA. As can be seen, the NMSE is not evenly distributed over all cepstral coefficients. While most coefficients suffer from the distortions at more or less the same level, c_3 and c_9 are much more severely affected. This uneven distribution may well explain

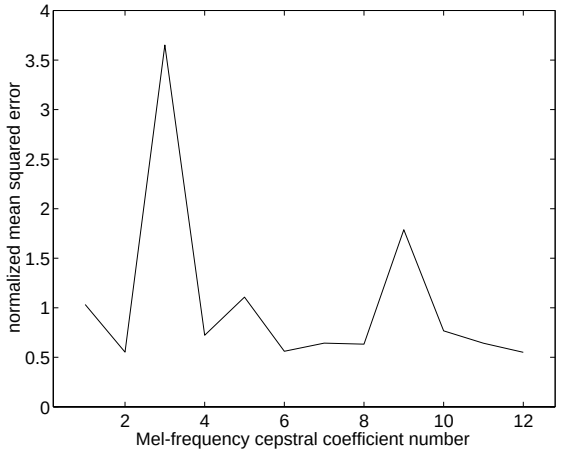

Fig. 7. Normalized mean squared error for the 12 static coefficients of the F2 features resulting from the SNR = 10 dBA distortion.

the WER reduction observed for our local distance function with acoustic backing-off. This type of robust local distance function effectively diminishes the impact of outliers (i.e., severely distorted feature values).

The data in Fig. 7 show that there is an interaction between the characteristics of the distortion and the way in which the statistical properties of the features are affected. This interaction may be quite complex. As a consequence, the fully smearing DCT may eventually give rise to features of which only a few are substantially distorted for a given type of noise. Informal experiments in which we tried to find a relation between NMSE and WER have not been successful. This suggests that we need to develop new measures that would allow us to rank different types of adverse conditions with respect to their impact on recognition performance. For these new measures, it is probably not enough to study only the specific manner in which a given noise source may affect a certain feature vector, like in the NMSE. It should be kept in mind that it is the distorted observation evaluated according to the current active state distributions that determines the best Viterbi path. Therefore, we would suggest a study of the contributions to the local distance for each individual feature component along (say) the N-best recognized paths. Such new measures would probably also allow one to select the best performing robustness strategy for a specific task.

The results of experiment 3 suggest that the effectiveness of acoustic backing-off in combination with P2 features is not limited to artificial distortions but can be generalized to more realistic conditions, provided that the noise is band-limited. Thus, our last hypothesis about the acoustic features and the robust local distance function appears to be supported by our experiments.

The results for babble and factory noise show that a robust local distance function in the form of acoustic backing-off cannot yield an improvement, at least not for the P2 features that were tested. An important difference between additive car noise on the one hand and additive babble and factory noise on the other is the proportion of frequency components that is still intact in the presence of noise. For car noise, which is band-limited, the proportion of undisturbed frequency components is larger than that for babble and factory noise, which are wide-band noises. This suggests that the effectiveness of acoustic backing-off (in combination with P2 features) is proportional to the number of raw feature components that is unaffected by the noise. This is a subject for further research.

In addition, it should be noted that distortions of the raw features cannot be considered to be local for wide-band noise. When distortions are not local to start with, the strategy to keep local distortions local may become questionable. Therefore, the P2 features might not have been the optimal feature representation to use in combination with wide-band noise. Additional research is required to test whether other feature representations are better suited when wide-band noise is present (e.g., representations in which distortions of many raw features are projected into a few components).

Despite its substantial advantage over a conventional local distance function for band-limited types of distortions, acoustic backing-off is not able to counteract the impact of this type of additive noise completely. Even under mild conditions, viz. band-limited stationary noise added at a $\mathrm{SNR} = 20$ dBA, performance degrades quite substantially (cf. Fig. 4). We assume that this finding can be explained in part by the fact that the impact of a distortion on the statistical distributions of the transformed parameters is difficult to predict, but that it is also quite conceivable that many types of distortions may result in small effects on all feature values. As a matter of fact, much of our thinking assumed some kind of equivalence between 'distortions' and 'outliers'. However, at least two qualitatively different kinds of distortions must be distinguished, viz. those which give rise to values which are (more or less) clear outliers and those which result in feature values falling well within the range of values observed during the training of the speech sound or state under consideration. Acoustic backing-off as an instantiation of MFT was designed to cope with feature values that are outliers with respect to the trained distributions. We have no reason to doubt that we have succeeded in reaching this goal, as testified by the

consistent performance gain for band-limited types of noise. However, it is not clear how acoustic backing-off might handle distortions that do affect feature values without pushing them widely beyond the center of gravity of the trained distributions. Of course, acoustic backing-off can be combined with other techniques that reduce the mismatch between clean and adverse conditions, like spectral subtraction. Yet, there remains a need to search for ways to handle the situation where almost all feature values are mildly corrupted. This will inevitably necessitate the exploration of ways to combine other sub-fields of robust pattern recognition with approaches related to MFT. One route that suggests itself is replacing the Gaussian mixtures by mixtures of distributions which are inherently more robust against large proportions of small errors.

At the same time, our results emphasize the need for a continued quest for a representation of speech signals that is inherently more robust against distortions. It is interesting to note that 'robustness' can be accomplished in two rather different ways. The first is by means of a representation that yields values in exactly the same distribution under clean and adverse conditions. This approach is well known and spectral subtraction may be viewed as the classic example of it. The second, somewhat less conventional, solution builds on the property of a technique like acoustic backing-off that outlier values can effectively be dealt with. With this in mind, the second solution would be a representation in which distorted raw feature values are transformed into a few outlier values. In such a scheme, the transform would be used to focus distorted raw feature values.

6. Conclusions

In this paper, two closely related points were studied that should improve probabilistic modeling and decoding for ASR in mismatched training-test conditions: the type of features used to represent the speech signals and the measure used to compute the distance between an observed feature vector and a previously trained model. As a starting point, one of the key ideas of MFT was adopted: distorted feature values should not be trusted in the same manner as clean observations (Cooke et al., 1996; Morris et al., 1998). It was discussed how this principle could be incorporated in an otherwise conventional ASR set-up by changing the local distance function. With respect to the acoustic features, it was argued that a distinction must be made between feature representations that smear spectrally local distortions over all feature vector components (F-type) and representations that limit smearing to a sub-set of the feature vector components used for modeling and recognition (P-type, cf. (Okawa et al., 1998)). This distinction is necessary to ensure that as few features as possible become corrupted. Next, a robust local distance function that is based on one of the principles of robust statistical pattern recognition, i.e., acoustic backing-off (de Veth et al., 1998, 2001) was investigated. By means of the new robust local distance function, it was ensured that corrupted features do not contribute significantly to the recognition decision.

In the context of connected digit recognition over the telephone, different ASR set-ups that were based on HMMs were evaluated. In all experiments, corrupted speech was created by adding noise to the original clean speech signals. The effects of low-range band-limited noise at SNRs of 20, 10 and 5 dBA and the effects of mid- and high-range band-limited noise at $\mathrm{SNR} = 10$ dBA were investigated in the first two experiments. In a third experiment, NOISEX car, babble and factory noise were added to the speech signals at an SNR of 10 dBA. The recognition performance for different combinations of feature type and method for local distance computation were compared.

The results for band-limited noise show that a partial-smearing transform is preferred over a full-smearing transform. Thus, for band-limited noise, it can be safely concluded that care must be taken to choose a feature representation in which possible noise sources affect as few feature vector components as possible. Moreover, the results indicate that a partial-smearing transform is preferred irrespective of the distance measure used (i.e., conventional or robust) as long as the noise is confined to a limited frequency range. Furthermore, the experimental evidence indicates that

acoustic backing-off appears to be effective in improving noise robustness for artificially created, band-limited noise as well as more realistic NOISEX car noise. For NOISEX babble and factory noise, however, acoustic backing-off is not capable of improving recognition performance, at least not in combination with the partial-smearing, within-vector filtered MFLECs. Therefore, it is concluded that the degree to which acoustic backing-off is effective appears to be dependent on both the feature type and the type of additive noise. We argue that this may be explained by the particular way in which the distortions are distributed over the different feature vector components: dependent on how distortions are distributed over the raw features (due to the noise characteristics) and how these distortions are redistributed over the actual acoustic features (due to the transformation used in the pre-processing stage), some components will be more heavily distorted than others. Acoustic backing-off can limit the impact of outliers, so that recognition is effectively based on those features that are least affected from a statistical point of view. For distorted features that are not outliers with respect to the distribution of feature values observed during training, acoustic backing-off cannot be expected to give an advantage.

Overall, the feature set based on within-vector filtered MFLECs in combination with acoustic backing-off consistently gave the best results for all band-limited noise conditions that were studied. For these features, acoustic backing-off reduced the WER by almost 40% with low-range, band-limited noise at SNR = 5 dBA added to the clean speech signals. For the same type of features, the robust local distance function reduced the WER by almost 50% for NOISEX car noise at SNR = 10 dBA.

Acknowledgements

The research of Johan de Veth was carried out within the framework of the Priority Programme Language and Speech Technology (TST). The TST-Programme is sponsored by NWO (Dutch Organisation for Scientific Research).

References

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120.

Cooke, M., Morris, A., Green, P., 1996. Recognising occluded speech. In: Proc. ESCA Workshop on the Auditory Basis of Speech Perception. Keele, UK, pp. 297–300.

Dautrich, B., Rabiner, L., Martin, T., 1983. The effects of selected signal processing techniques on the performance of a filter bank based isolated word recognizer. Bell System Tech. J. 62 (5), 1311–1336.

den Os, E., Boogaart, T., Boves, L., Klabbers, E., 1995. The Dutch Polyphone corpus. In: Proc. Eurospeech. pp. 825–828.

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

de Veth, J., Cranen, B., Boves, L., 1998. Acoustic backing-off in the local distance computation for robust automatic speech recognition. In: Proc. Internat. Conf. Spoken Language Process. pp. 1427–1430.

de Veth, J., Cranen, B., Boves, L., 1999. Acoustic backing-off as an implementation of missing feature theory. Internal report, Priority Programme Language & Speech Technology, no. 81. Also: http://lands.let.kun.nl/literature/deveth.1999.2.html.

de Veth, J., Cranen, B., Boves, L., 2001. Acoustic backing-off as an implementation of missing feature theory. Speech Communication 34 (3).

Dupont, S., Bourlard, H., Ris, C., 1997. Robust speech recognition based on multi-stream features. In: Proc. ESCA–NATO Workshop on Robust Speech Recognition for Unknown Communication Channels. Pont-a-Mousson, France, pp. 95–98.

Gales, M., 1998. Predictive model-based compensation schemes for robust speech recognition. Speech Communication 25, 49–75.

Hassall, J., Zaveri, K., 1979. Acoustic Noise Measurements. Brüel & Kjær, Denmark.

Huber, P., 1981. Robust Statistics. Wiley, New York.

Huerta, J., Stern, R., 1998. Speech recognition from GSM codec parameters. In: CD-ROM of Proc. Internat. Conf. Spoken Language Process.

Hunt, M., Bateman, D., Richardson, S., Piau, P., 1991. An investigation of PLP and IMELDA acoustic representation and of their potential for combination. In: Proc. Internat. Conf. Acoust. Signal. Speech Process. pp. 881–884.

Kharin, Y., 1996. Robustness in Statistical Pattern Recognition. Kluwer, Dordrecht.

Lee, C.-H., 1998. On stochastic feature and model compensation approaches to robust speech recognition. Speech Communication 25, 29–47.

Lee, C.-H., Huo, Q., 1999. Adaptive classification and decision strategies for robust speech recognition. In: Proc. Workshop on Robust Methods for ASR in Adverse Conditions. Tampere, pp. 45–52.

Lippmann, R., 1997. Speech recognition by machines and humans. Speech Communication 22, 1–15.

Lippmann, R., Carlson, B., 1997. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In: Proc. Eurospeech. pp. 37–40.

Morris, A., Cooke, M., Green, P., 1998. Some solutions to the missing feature problem in data classification, with applications to noise robust ASR. In: Proc. Internat. Conf. Acoust. Signal Speech Process. pp. 737–740.

Nadeu, C., Hernando, J., Gorricho, M., 1995. On the decorrelation of filter-bank energies in speech recognition. In: Proc. Eurospeech. pp. 1381–1384.

Ney, H., 1999. Summary: Speech recognition – Where do we stand? In: Eurospeech 1999. http://tel.ttt.bme.hu/Eurospeech99/data/keyney.zip.

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

Okawa, S., Bocchieri, E., Potamianos, A., 1998. Multi-band speech recognition in noisy environments. In: Proc. Internat. Conf. Acoust. Signal Speech Process. pp. 641–644.

Tibrewala, S., Hermansky, H., 1997. Sub-band based recognition of noisy speech. In: Proc. Internat. Conf. Acoust. Signal Speech Process. pp. 1255–1258.

Vizinho, A., Green, P., Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study. In: Proc. Eurospeech. pp. 2407–2410.

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