State dependent feature component selection for noise robust ASR

Bert Cranen & Johan de Veth

Radboud University Nijmegen, The Netherlands
{B.Cranen, J.deVeth}@let.kun.nl

Abstract

The acoustic environment in which speech is recorded has a strong influence on the statistical distributions of observed acoustic features. In order to make ASR insensitive to noise it is crucial that these distributions are similar in the training and testing condition. Mostly, it is attempted to compensate for the impact of noise by estimating the noise characteristics from the signal. In this paper we explore the feasibility of a new method to increase noise robustness: We try to exploit a priori knowledge stored in clean speech models. Using Mel bank log-energy features, recognition is done by ignoring the model components for features that contained little energy during training. This strategy aims at recognition results that are determined more strongly by the match in the high-energy rather than by the mismatch in the low-energy model components. Application of the new method to clean speech data confirms that discarding components below a certain energy threshold does not deteriorate recognition performance. Experiments with noisy data, however, show that performance gains are relatively small. This paper explains why that is the case and why, despite the limited success, the outcomes suggest that the method still could prove to be a valuable addition to data-driven methods like (bounded) marginalisation.

1. Introduction

As long as speech is produced in an acoustic environment in which only a mild and stationary background noise is present, the short-term spectro-temporal energy distributions that are associated with the speech can be safely considered as drawn from the same population. Consequently, all properties of the incoming signal are probably equally useful for decoding the speech signal by means of an ASR engine based on hidden Markov models (HMMs). However, if the type and level of the background noise is not known beforehand and fluctuates over time, HMM based speech recognizers trained with speech from a quiet environment usually do not perform well. This phenomenon is easy to understand: Most of the spectro-temporal regions in which the clean speech sounds show relatively little energy are flooded with energy from the background sources. As a result, the distributions of the acoustic features that had low energy in the absence of noise may differ substantially from those of noisy speech, causing the distance between competing hypotheses to become strongly dependent on the background noise. In order to limit the number of recognition errors that are caused by the noise, methods must be deployed to diminish this training-test mismatch.

In the past decades, many different techniques have been developed to make ASR less sensitive to environmental noise. Although the implementation details differ vastly, these techniques all aim for the same thing: Making the distance between the acoustic feature probability density functions (pdfs) of clean and noisy speech as small as possible, and as insensitive to the noise condition as possible. Regardless whether one looks at methods that try to compensate for the noise by adaptations at the feature level (e.g., time domain noise reduction [1] or spectral subtraction [2]), or at methods that directly try to affect the feature distributions themselves (e.g., histogram normalisation [3] or model compensation [4]), all these methods have in common that the required compensation is directly and exclusively estimated from the noisy input. Also in missing feature approaches [5], the noisy input itself is taken as a starting point for selection of reliable and unreliable features.

Given the fact that it is the rule rather than the exception that an unknown signal consists of a mixture of sounds, it is hardly surprising that making a distinction between acoustic features that are either dominated by the speech signal (reliable) or by the background noise (unreliable) helps greatly to improve recognition performance. In practice, however, the success of a missing feature approach is strongly determined by the degree with which features can be correctly labelled as reliable or unreliable. Using solely data-driven estimation techniques, it appears quite difficult to avoid labelling errors [6].

It is intuitively plausible to assume that labelling features as reliable or unreliable is easier in a domain where they have a straightforward physical interpretation, such as the spectro-temporal energy domain. At the same time, it is true that filter band energy features have the drawback that they are strongly correlated, which makes it more difficult to build compact statistical models. Yet, we decided to use filter band energies throughout this paper, because we want to investigate whether it is possible to use a straightforward physical interpretation as the starting point for a novel technique for handling potentially corrupted acoustic features.

Whereas ASR engines cannot deal well with features that are erroneously labelled as reliable, humans are extremely flexible in ignoring information that does not fit the pattern they are looking for. Several studies

suggest that the superior recognition performance of humans can at least partly be attributed to their capability to rely on expectations about what is likely to be perceived [7] [8] [9]. In this paper, we explore to what extent ASR can be made more noise robust by using a priori knowledge that has been acquired during acoustic model training on clean speech. We apply an energy criterion for selecting a subset of clean speech model components that is deemed robust enough to be used at recognition time, and we set up a matching scheme that adapts the distance measure to the model for which the likelihood of the signal features is computed. In this manner, we aim at diminishing the impact of features whose values, according to a specific hypothesis, are corrupted by noise. In fact, our approach boils down to emphasizing the match between models and unknown signals in the high-energy spectro-temporal regions, rather than the mismatch in its low-energy regions. One of the potentially adverse consequences of this approach is that it also ignores matches in low-energy regions of the spectro-temporal plane.

The feasibility of this new approach was tested by means of a series digit recognition experiments. First, we used clean speech in order to investigate to what extent information about low-energy features in the clean speech models can be discarded without harming recognition performance. The results of this experiment provided insight into the relative importance of features with high and low energy. It also gave an upper bound estimate for the recognition performance one could hope for when the same strategy is applied to noisy speech. Next, we applied our technique to speech with additive babble noise. The results of this experiment were somewhat disappointing. Therefore, we designed and conducted a number of follow-up experiments to establish whether the lack of success is due to decisions that had to be made in the implementation of the technique (and thus can be changed for the better), or whether the causes are more fundamental (and therefore cannot easily be remedied).

The rest of the paper is organised as follows. In Section 2 we describe our novel method for scoring local distance and the general experimental set-up. More in particular in Section 2.2 we describe a new distance measure which may differ for each and every state. In Section 3 we present the results of the digit recognition experiment both for clean and noisy data and we discuss the limitations of the proposed method. In Section 3 we report on three experiments that were conducted to investigate three factors that might explain the disappointing results for noise robustness. Finally, in Section 4 we summarize the most important findings and draw conclusions.

2. Method

2.1. Using expectations to ignore corrupted features

Research on human hearing has shown that spectro-temporal signal components with high intensity dominate the neural response [9]. To remain in a domain where perceptual relevance can be inferred directly from the magnitude of the features (at least to a first order approximation), we use Mel filter band spectra throughout this paper. The output of the filter bands are expressed in terms of log-magnitude. By taking the logarithm, we ensure that spectro-temporal regions with high speech energy are less affected by background noise than regions with low speech energy [1]. In other words, one may not only expect that the high-energy portions of clean speech models contain most of the relevant information, but also that they are more reliable for decoding purposes.

In conventional ASR systems, decoding is done by evaluating the likelihood of an unknown speech vector using *all* its acoustic features for comparison with the statistical distributions of the the corresponding features observed during training. As long as the spectro-temporal regions with low speech energy (that are easily polluted by noise) can be correctly recognized as non-reliable speech information, marginalization approaches can ensure that these regions do not harm the decoding decision. In practice, however, it appears very difficult to correctly distinguish reliable and unreliable features without having information about the underlying speech signal [5] [6].

In order to investigate to what extent information about features that exhibited low energy during training is useful for reducing the impact of high-energy features that have been erroneously associated with speech, we modify the distance computation in a model dependent way. Taking the speech models as a reference, as opposed to the unknown speech signal, we ensure that a high observation value in a certain frequency band is not going to be considered as reliable information by any speech model for which training has learned to expect a low value in that band. Thus we strive to obtain a matching procedure that focuses on the more important and reliable spectro-temporal regions and that is affected as little as possible by the presence of noise in regions that are known beforehand to be unimportant for the decoding of a specific model. In the next section we explain in detail how the distance computation was modified.

2.2. State dependent feature selection

In analogy to the missing feature approach [6], we split the components of each acoustic observation vector ($\vec{x}$) in two subsets. The first subset consists of features that must fit the model pdf as closely as possible because this information is considered mandatory for a reliable recognition ($\vec{x}_R$). The other subset will not be considered at all during the match assuming that these components have too big a chance to represent noise ($\vec{x}_N$). Note that in contrast to the standard missing feature approach, the subsets are chosen differently for different HMM states. This means that both $\vec{x}_R$ and $\vec{x}_N$ are functions of the hypothesized state s_j: $\vec{x}_{R(s_j)}$ and $\vec{x}_{N(s_j)}$. The posterior probability for

---

[1]Denoting the energy of the desired signal by E_X and the energy of the (uncorrelated) noise by E_N one can write: $\log(E_X + E_N) = \log(E_X) + \log(1 + E_N/E_X) \approx \log(E_X)$ if $E_N \ll E_X$.

a given state s_j can be written as:

$$P(s_j|\vec{x})$$
$$= P(s_j|\vec{x}_{R(s_j)}, \vec{x}_{N(s_j)})$$
$$= \frac{P(\vec{x}_{R(s_j)}, \vec{x}_{N(s_j)}|s_j) \cdot P(s_j)}{P(\vec{x}_{R(s_j)}, \vec{x}_{N(s_j)})}$$
$$= \frac{P(\vec{x}_{R(s_j)}|s_j) \cdot P(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)}, s_j) \cdot P(s_j)}{P(\vec{x}_{R(s_j)}) \cdot P(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)})}$$
$$= P(s_j|\vec{x}_{R(s_j)}) \cdot \frac{P(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)}, s_j)}{P(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)})} \quad (1)$$

We assume that the relevant parts $\vec{x}_R$ and the non-relevant parts $\vec{x}_N$ of the feature vector are mutually exclusive. The term $p(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)}, s_j)$ represents the feature distributions for all vector components that are assumed unimportant for encoding the target speech information in state s_j. This pdf describes the distributions of features that represent the valleys in the spectrum, i.e., for clean speech the characteristics of silence and for noisy speech the spectral characteristics of the background noise. The term in the denominator $p(\vec{x}_{N(s_j)}|\vec{x}_{R(s_j)})$ represents a similar distribution, i.e. the distribution of features that are considered unimportant for the current state, but averaged over all possible states. Thus, this pdf includes both speech and non-speech related features. As a first approximation we will assume that the quotient in Eq. (1) only contains disinformation and may better be ignored. Thus, only feature vector components that are marked as relevant will be used for decoding, by using

$$P(s_j|\vec{x}) = \frac{P(\vec{x}_{R(s_j)}|s_j) \cdot P(s_j)}{P(\vec{x}_{R(s_j)})} \quad (2)$$

Note that evaluation of this probability only requires a priori knowledge. The identity of the vector components is entirely determined by the models. No processing of the unknown signal is involved to determine which features are expected to represent reliable speech information or not.

Finding the optimal {frame,state}-path can be done in the usual way with dynamic programming. With $D_{t,j}$ denoting the cumulative distance, representing the minimal cost associated with ending in state s_j at time t (assuming that state s_i was visited at time $t-1$), having observed the series of acoustic vectors $\vec{x}(1), \vec{x}(2)...\vec{x}(t)$, we can write:

$$D_{t,j}$$
$$= \min_i \{D_{t-1,i} - \log[p(s_j(t)|s_i(t-1)])\} +$$
$$- \log\left[\frac{P(\vec{x}_{R(s_j)}(t)|s_j) \cdot P(s_j)}{P(\vec{x}_{R(s_j)}(t))}\right] \quad (3)$$

Since calculation of $P(\vec{x}_{R(s_j)}|s_j)$ and $P(\vec{x}_{R(s_j)})$ only involves a subset of the features, the components in these terms may be different for each hypothesis to be evaluated. As a consequence, the denominator term $P(\vec{x}_{R(s_j)})$ cannot be factored out in the usual way when comparing different alternatives and needs to be explicitly estimated for each individual state. Interpreting this term as

the prior probability of an observation, it can be estimated by calculating the overall feature distribution of all available training data and making the desired selection.

2.3. Digit recognition experiments

2.3.1. Speech material
The speech material for our experiments consisted of connected digit strings and was taken from the Dutch POLYPHONE corpus [10]. From this corpus, comprising speech that has been recorded over the public switched telephone network in the Netherlands, we selected connected digit strings with 3 to 16 digits per string. 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. [11]) 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.

For recognition experiments with noisy data, NOISEX babble noise was added to the clean speech signals resulting in signal-to-noise ratios (SNRs) of 15, 10, and 5 dBA. Care was taken that the amplitude of the speech was not changed when adding the noise. More details about the speech material can be found in [12].

2.3.2. Acoustic pre-processing
From the 8 kHz sampled speech signal 16 Mel-frequency log-energy coefficients (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, the 16 filter band energy values were calculated, with the filter bands triangularly shaped and uniformly distributed on a Mel-frequency scale (covering 122.0-2143.6 Mel; this corresponds to the linear range of 80-4000 Hz). In addition to the 16 MFLECs, we also computed the total log-energy for each frame. These signal processing steps were performed using HTK [13]. The 17 static coefficients were augmented with (smoothed) first-and second-order time derivatives (delta- and delta-delta-coefficients) to arrive at 51-dimensional feature vectors.

2.3.3. HMMs
The ten Dutch digits were represented as 10 whole-word models. The number of states in each model was chosen proportional to the number of phones in the word. Furthermore, we used three additional three-state models for silence, background noises and out-of-vocabulary speech. Each unit was represented as a left-to-right HMM. For these models the total number of states was 108 (99 for the words plus 9 for the silence

and noise models). We used HTK for training and testing HMMs [13]. To determine the optimal number of Baum-Welch iterations, we followed the cross-validation scheme described in [11]. The initial single-Gaussian models were split up to four times, resulting in recognition systems with 2, 4, 8, and 16 Gaussians per state. Thus, the most complex model set contained 1728 Gaussians (with diagonal covariance matrices) in total. The models were trained using clean speech only. The recognition syntax used during cross-validation and testing was defined so that connected digit strings of 3 to 16 digits could be recognized.

Gaussian mixture HMMs are computationally inconvenient if one desires to identify and modify the contribution of a specific vector component to the total cost of a {frame,state}-path. Therefore we used a work-around by reshaping each N-Gaussian mixture state into a set of N parallel single-Gaussian states. The transition probabilities to separate parallel states were determined by the original mixture weights. Using a maximum likelihood decoding, this conversion ensures that each vector component makes an independent contribution to the local distance because its likelihood is always evaluated against a single-Gaussian pdf. After conversion to a topology with parallel states, the original model sets of 108 states with 8 and 16 Gaussians per state, respectively, were converted into model sets with 864 and 1728 single-Gaussian states. Experiments indicated that the recognition performances of the original and converted model sets did not differ significantly.

For computation of the denominator term in Eq. (2), a single-Gaussian single-state HMM was trained using only the speech portions in the recordings of the entire training set.

2.3.4. State dependent component selection

The following procedure was used to select the components in each single-Gaussian HMM state, which, according to the expected energy value, was judged robust enough to be included in the distance computation. First, an absolute energy level was chosen (in the linear frequency domain) above which an observation was considered to represent relevant speech energy. Because the first 16 static coefficients in our acoustic vectors represent log-energy values computed on a Mel-scale, the chosen absolute energy level was converted to a corresponding Mel-energy threshold value for each Mel-frequency band used. For the total log-energy, the 17-th static coefficient in our acoustic vectors, a similar procedure was followed. Next, whenever s_j was hypothesized during decoding, a component in a single Gaussian density was assigned to $\vec{x}_{N(s_j)}$ (i.e., being irrelevant) when 95% of the trained probability density mass fell below the threshold. Otherwise, that particular component was assigned to $\vec{x}_{R(s_j)}$ (i.e., being relevant for decoding purposes)[2]. A gradual

---

[2]Note that this procedure differs from the one used in [14]. We prefer the current one, because the use of an absolute energy level allows us to maintain a closer link to the physics of the recorded signals.
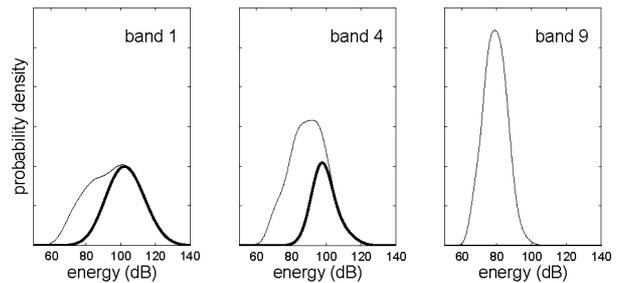

Figure 1: Thin lines: Original pdfs of 8-Gaussian mixtures for energy bands 1, 4, and 9; Thick lines: Effective pdf (the sum of 8 parallel single-Gaussians) after discarding all low-energy components of which 95% of the pdf probability mass lies below a pdf threshold of 100 dB.

increase of the chosen absolute energy level allowed us to control the number of low-energy model components that were excluded from the distance computation.

It is impossible to apply a separate amplitude criterion to mark delta and delta-delta components as having a higher or lower a priori likelihood to be relevant: Values for deltas and delta-deltas can both be positive and negative, and, in contrast to the static energy values, there is no law from physics that allows to predict their vulnerability to noise. Therefore, we decided to mark a delta component in a Gaussian distribution as (ir)relevant whenever the corresponding static component was marked as such. This rule was also applied to the delta-deltas.

The effect of our strategy to mark model components as relevant or irrelevant is illustrated in Fig. 1. The original pdf of three different feature components in the same multi-variate 8-Gaussian mixture HMM state is shown as a thin line. The estimated effective pdf after selection (obtained by summing only those components that were retained in the corresponding 8 parallel states) is shown as a thick line. The energy threshold applied for selection was 100 dB for each static component. In the rest of this paper we will refer to this threshold as *pdf threshold*. As can be observed from Fig. 1, the degree to which components are marked as relevant (i.e., the degree to which they should be effectively used during recognition) is different for each component. Above the pdf threshold (in this example 100 dB) the original and the new pdf overlap completely. In the left-most panel, corresponding to energy band 1, more components are retained than in the middle panel, corresponding to energy band 5. For energy band 9 (right-most panel) none of the original model components are retained in the distribution, meaning that observation values of band 9 are ignored altogether during the distance computation whenever this particular state is hypothesized.

As becomes obvious from Fig. 1, the degree of control over how the low-end tails of the pdfs are effected, is rather limited: The order in which different components are discarded is determined by the accidental combination of mean and standard deviations of the Gaussian components in the original models. Furthermore, one

should be aware that masking out model components below the pdf threshold in one or more of the parallel single-Gaussian states does not imply that the set of N parallel states replacing the original N-Gaussian mixture model cannot 'feel' observation data at all. As a consequence, the competition between different models may change in a rather random fashion for different pdf thresholds, especially when an observation vector contains a large proportion of low-energy values.

3. Results and discussion

3.1. Discarding low-energy model components

In a first series of experiments, recognition performance was studied for clean and noisy data as a function of the pdf threshold. All states in all models were treated similarly: Whenever a component occurred of which 95% of the probability mass was lying below the chosen pdf threshold, it was discarded (cf. Section 2.3.4). The recognition accuracies obtained are shown in Fig. 2 for model sets with 864 and 1728 Gaussians, respectively. The horizontal dashed lines in this, as well as in all subsequent figures, indicate the recognition accuracy for the conventional ASR system in which all model components were retained.

Figure 2 shows that, for clean speech, increasing the pdf threshold causes the recognition accuracy to decrease. For pdf threshold values up to 100 dB the decrease is so gradual that the sub-threshold model components can be discarded without seriously falling below baseline recognition performance. For the model set with 864 Gaussians, 24.9% of the model components are discarded when the pdf threshold is set to 100 dB. For the model with 1728 Gaussians, 27.3% of the model components are discarded using a pdf threshold of 100 dB.

For noisy speech, the recognition performance improves slightly for pdf thresholds up to 100 dB. Above this value, recognition accuracy starts to deteriorate again and soon drops well below the baseline performance. For example, for 10 dBA noisy speech, the accuracy at a pdf threshold value of 100 dB improves from 65.7% to 69.7% for models with 864 Gaussians and from 67.6% to 70.2% for models with 1728 Gaussians.

The results in Fig. 2 indicate that the effectiveness of the state dependent component selection method is rather limited. Apparently, ignoring the training-test mismatch caused by those features that appeared irrelevant for the decoding of clean speech only helps slightly to achieve a better recognition in noisy conditions. The fact that the improvement is so little, suggests that there is still a huge amount training-test mismatch left in the features that are not ignored. In order to obtain a clearer view on this matter, a series of follow-up experiments was conducted.

3.2. Using additional information from the signal

As already discussed at the end of section 2.3.4, the manner in which the component selection takes place causes the shapes of the low-end tails of the effective pdfs to
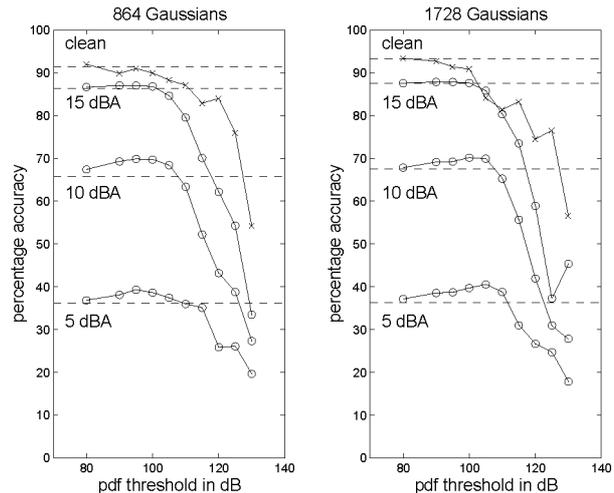

Figure 2: Recognition accuracy as a function of the pdf threshold used to discard model components. From top to bottom the curves represent clean speech, and speech with babble noise at SNRs of 15, 10, and 5 dBA. Left panel: model set with 864 Gaussians; Right panel: model set with 1728 Gaussians. Dashed lines indicate recognition performance with all model components retained.

change quite coarsely. These changes are likely to affect the competition between hypothesized models in a rather uncontrolled way. We conjectured that this effect might be one of the factors limiting the effectiveness of the state dependent component selection method. Therefore, we modified our decoding scheme in such a way that we had better control over the extent to which observations falling in the tail left of the pdf threshold affected the recognition result. We let an observation contribute to the posterior probability in Eq. 2 only when both the model component for the hypothesized state was considered relevant (according to the pdf threshold criterion) and the corresponding feature in the observation vector was lying above a second pre-defined energy threshold (that we will from now on denote by *observation threshold*). Note that using an observation threshold (as long as no pdf-threshold is used) is equivalent to a conventional observation-driven marginalisation approach [6] with a rather crude criterion to establish the reliability of a feature.

Using only the model set with 864 Gaussians, tests with this modified approach were run for clean and 10 dBA noisy speech. The results of these experiments are shown in Fig. 3 for two conditions: (1) when no state dependent component selection was applied (left panel) and (2) for a pdf threshold of 100 dB (right panel). We shall first focus on the results for noisy speech. The results for clean speech will be discussed later.

As can be inferred from Fig. 3, application of an additional observation threshold does not improve recognition performance in the noisy condition. This holds regardless whether state dependent component selection is used or not. The difference between the accuracy levels
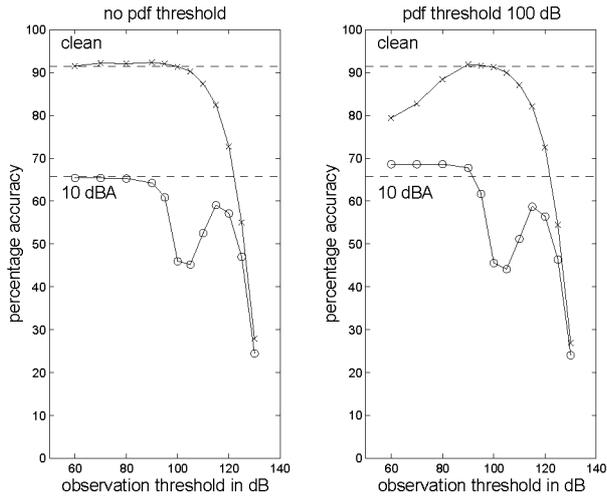
Figure 3: Recognition accuracy as a function of observation threshold. Left panel: No model component selection. Right panel: Model components selected using a 100 dB pdf threshold. Upper curves: Clean speech; lower curves: 10 dBA noisy speech. Dashed lines: Accuracy with all model components retained.

Figure 4: Similar results as in Fig. 3, but with observation thresholds chosen with respect to the average noise spectrum.

in the left and right panel for observations thresholds below 90 dB, is the same as observed in Figure 2. The fact that observation thresholds up to 90 dB do not give rise to any changes in recognition accuracy suggests that the majority of the noisy test observations have values larger than 90 dB. This was confirmed by looking at the feature distributions of the noisy test data. Because hardly any observations lie in the range where the left tails of the effective pdfs may play a role, it is not surprising that a possible adverse effect due to an uncontrolled behaviour of these slopes, does not become manifest in the noisy data experiments.

Note that the curves in both panels are virtually identical above 95 dB. Of course, this must be expected since in that regions no feature values < 95 dB occur and because the effective pdfs above 100 dB are indistinguishable from the original pdfs. In other words, the recognizers in the left and right panel are virtually identical. Inspection of the recognition results revealed that the first part of the dip (observation thresholds of 90-105 dB) corresponds with a decrease of deletion errors and an accompanying (larger) increase of substitution and insertion errors. For observation threshold larger than 105 dB, the balance reverses: deletions start to increase, and substitutions and insertions decrease. We interpret this changed balance as the combined result of a positive effect due to the removal of noise energy, which is counteracted by the negative effect of removing speech energy. The net effect being negative, is a clear indication that using a single, fixed energy threshold in the linear frequency domain for deriving data-dependent reliability masks is too crude a method. Obviously, a change in observation threshold does affect which model wins the competition. However, it can not ensure that the correct model wins more often.
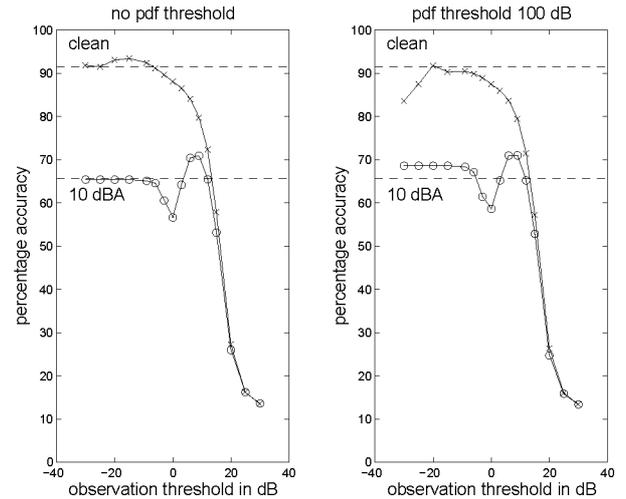
3.3. Using average noise characteristics

To find out whether using a less crude criterion for estimating the data-dependent reliability mask would change the results in a qualitative way, we subsequently applied observation thresholds that were chosen relative to the mean energy level of the noise in each band. Because we used the same fixed observation thresholds for each test utterance, this procedure might still be far from optimal. Nevertheless, it ensures that a larger proportion of the features that are dominated by speech (c.q. noise) energy are marked correctly as reliable (c.q. unreliable). The results for clean and 10 dBA noisy speech are shown in Fig. 4 for a condition in which no state dependent component selection is applied (left panel) and for a pdf threshold of 100 dB (right panel). As expected, this noise-dependent reliability mask does a better job than the previous one: At an observation threshold level between 6-9 dB relative to the mean noise energy (i.e. where the major part of the noisy features has been ignored in the distance computation) a maximum accuracy of 71.0% is obtained. This is slightly better than the performance for state dependent component selection alone (68.6%), but still far off from clean speech performance at that point (79.5%). No additional accuracy gain is observed when state dependent component selection is combined with data-driven feature selection. This suggests that, just like in section 3.2, for observation thresholds > 3 dB the recognition is fully determined by data-driven feature selection process.

3.4. Clean speech results

Now, let us return to the discussion of the recognition results for clean speech. All results for the clean data obtained in the experiments discussed so far, show that increasing energy thresholds (pdf threshold, observation threshold, or both) above 100 dB always leads to loss of recognition performance. Inspection of the feature dis-
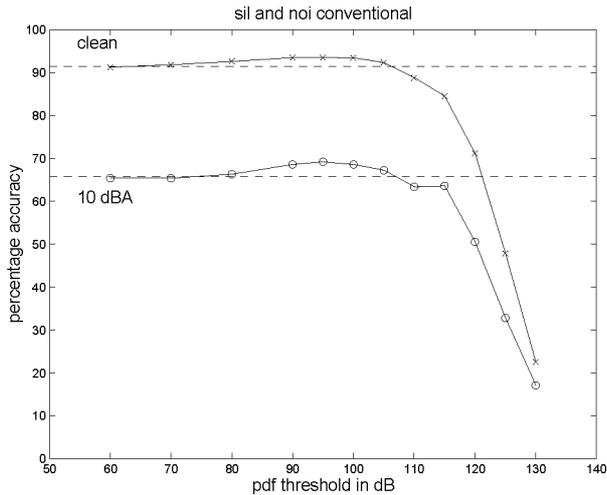
Figure 5: Recognition performance as a function of the energy level used to select model components for the model set with 864 Gaussians. Component selection was applied to all HMM units, except those representing silence and non-speech noise. Upper curve: clean speech; lower curve: 10 dBA babble noise. Dashed lines indicate recognition performance with all model components retained.

tributions of the speech parts of the clean training data showed that the modes of the distributions lie in the range of approximately 125 dB for the first five energy bands, to about 100 dB for energy bands 6−16. Our finding that the full dynamic range above 100 dB is required to maintain recognition accuracy thus implies that information from all filter banks must be used. The improved results for the noise related observation thresholds (cf. section 3.3) can be interpreted that some sort of local SNR criterion is able to preserve more of the high frequency features. This suggests that it might be better to base the model component selection on the mean speech energy per band in the training database instead of an absolute energy criterion as in section 3.1.

A number of peculiarities with respect to the recognition results of clean speech have not been discussed yet. First, in Fig. 2 the accuracy for clean speech degrades steadily, while for the noisy conditions the recognition performance either does not change or improves slightly. Second, both in Figures 3 and 4, it can be observed that the combination of a pdf threshold of 100 dB and a variable observation threshold gives rise to a performance drop for very low values of the observation threshold. We hypothesized that the latter effect is due to an extra, un-intended model-data mismatch. Since some models contain more low-energy components than others, the effective pdfs of different models are affected to a different degree when the lowest energy components are ignored. This is likely to disturb the competition between models, particularly because we treated all models similarly. The rationale behind ignoring spectral components with low energy was that these were unimportant for conveying in-

formation about the identity of speech sounds. Of course, this reasoning does not apply to silence and background noise, while especially the effective pdf of these models are likely to change drastically as a function of the pdf threshold. We therefore conducted an additional experiment in which state dependent component selection was applied to the speech models only and not to the silence and noise model. The results are shown in Fig. 5.

First, Fig. 5 shows that the recognition performance in the clean condition does no longer degrade (as in Fig. 2), but remains above the baseline level up to a pdf threshold of about 105 dB. For the noisy condition, the improvements are primarily found at pdf threshold levels greater than 110 dB. As expected, considering all non-speech model components as informative instead of discarding the ones with low-energy reduces the mismatch.

4. Summary and conclusions

Acoustic models trained with clean speech, contain information about how much energy must be expected in different frequency bands for each HMM state. Since the valleys in log-magnitude speech spectra are more easily dominated by non-speech energy than the peaks, one might expect that low-energy bands are relatively unimportant for recognition. We hypothesized that it might be profitable to use a priori knowledge about low-energy areas to decide for each model independently which components are likely to be unreliable and should better be ignored. To be able to test this hypothesis, we implemented a state dependent component selection method: During recognition the contributions to the local distance of the low-energy components from the clean speech models are ignored irrespective of the observed feature value. Thus, we obtained a matching procedure that emphasizes the match in the high-energy regions and ignores mismatch in the regions that, according to the clean speech models, are expected to have low energy.

We tested our approach for a digit recognition task. Using a fixed energy threshold (in the linear frequency domain), we first selected the pdf-components in each state of all our *models* that were below this threshold and discarded their contribution to the local distance computation. In addition, we experimented with a marginalisation approach in which we discarded the distance contributions of all *observations* that were below a certain threshold. For clean speech, we found that recognition accuracy could be maintained only if we allowed the features to span a dynamic range of at least 30 dB.

Using models containing only those components that appeared to be indispensable for recognition of clean speech, we found that recognition performance for noisy speech was improved, but only slightly (for speech with babble noise at an SNR of 10 dBA recognition accuracy improved by approximately 3%, i.e., from 65.3% to 68.6%). Just like with clean speech, we found that attempts to reduce the impact of the noise further by ignoring more of the low-energy features were counterproductive. Simply discarding information of features with

an energy in the low end of the 30 dB dynamic range has an adverse effect: Together with the noise, also important speech information is discarded at the same time.

This finding severely limits the applicability of the idea to use an absolute energy criterion in combination with filter bank features to select robust features on a priori grounds alone. The state dependent component selection method is helpful in ignoring features of which it is clear beforehand that they are unimportant for the recognition of clean speech. Unfortunately, however, these features appeared to be responsible for only a minor part of the recognition errors in our tests. The fact that a relatively large proportion of the model components that are indispensable for maintaining recognition accuracy have relatively low energy, makes that even at very moderate noise levels the majority of the features are affected so that training-test mismatch is hardly reduced. Within the decoding scheme we used, our silent assumption that features that are robust against noise would probably also be the most important ones from an information coding point of view is obviously wrong.

Since marginalisation techniques use the unknown *input data* to estimate which features are reliable or not, while our state dependent component selection uses *a priori knowledge from clean speech models*, the two methods are to some extent complementary. This suggests that state dependent component selection constitutes a method that might have potential value, but only if it used in combination with other methods to improve noise robustness, e.g., (bounded) marginalisation techniques [5] [6]. We expect that in combination with more sophisticated data-driven approaches like in [6], a state dependent component selection approach allows to recover from part of the errors that blind estimation techniques are bound to make. After all, it is likely that at least part of the features that are erroneously labelled as reliable will be ignored by a state dependent component selection mechanism. The current study does not give any clear indications whether such an added value really exists and more research would be needed to verify this.

Another important conclusion concerns our finding that speech and non-speech models should be treated differently. It was shown that discarding the lowest energy components affects the left slope of the effective pdf that replaces the original N-Gaussian mixture pdf of a state. As expected, altering these slopes has a strong effect on the competition between the models that are hypothesized during the search. Many extra recognition errors may result from favouring the wrong models. For non-speech models we concluded that it is better to refrain from state dependent component selection and to use all available features. Using the state dependent component selection mechanism for speech models only, we found that, both for clean speech and noisy speech, higher energy thresholds could be used (discarding more model components) without decreasing recognition performance. This is an interesting finding to pursue. The more more components can be discarded, the

more promising the mechanism of state dependent component selection becomes as a potentially valuable addition to traditional marginalisation methods. For similar reasons, it is also interesting to speculate about what will happen to the sparsity of the state representations if one changes the criterion for discarding model components. For instance, we would expect that using thresholds derived from the mean speech energy per filter band as seen during training, rather than using absolute thresholds, will also allow to discard more components without lowering the recognition accuracy. Additional experiments are needed to show the true potential of state dependent component selection.

5. References

[1] Noé, B. Sienel, J., Jouvet, D. Mauuary, L. Boves, L. de Veth, J. and de Wet, F. "Noise reduction for noise robust feature extraction for distributed speech recongnition", Proceedings of Eurospeech 2001, Aalborg, Denmark, 433-436, 2001.

[2] Boll, S.F. "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-27(2):113–120, 1979.

[3] Hilger, F and Ney, H. "Quantile based histogram equalisation", Proceedings of Eurospeech 2001, Aalborg, Denmark, 1135–1138, 2001.

[4] Lee, C.H "On stochastic feature and model compensation approaches to robust speech recognition", Speech Communication, 25:29-47, 1998.

[5] Barker, J. Cooke, M. and Green, P. "Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise", Proceedings of Eurospeech 2001, Aalborg, Denmark, 213–216, 2001.

[6] Cooke, M. Green, P., Josifovski, L. and Vizinho, A.,"Robust automatic speech recognition with missing and unreliable data", Speech Communication, 34:267–285, 2001.

[7] Engel, A.K. Fries, P. and Singer, W. "Dynamic predictions: oscillations and synchrony in top-down processing", Nature Reviews Neuroscience, 2, 704–716, 2001.

[8] Bajcsy, R. "Active perception", Proceedings of the IEEE, 76(8):996–1005, 1988.

[9] Moore, B. An introduction to the Psychology of Hearing, Academic Press, New York, 1997.

[10] den Os, E., Boogaart, T., Boves, L. and Klabbers, E. "The Dutch Polyphone corpus", Proceedings of Eurospeech 1995, pp. 825-828, 1995.

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

[12] de Veth, J., de Wet, F., Cranen, B. and Boves, L. "Acoustic features and a distance measure that reduce the impact of training-test mismatch is ASR", Speech Communication, 34, 57–74, 2001.

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

[14] Cranen, B., and de Veth, J. "Active perception: Using a priori knowledge from clean speech models to ignore nontarget features", to appear in Proceedings of ICSLP 2004.