ACOUSTIC BACKING-OFF AS AN IMPLEMENTATION OF MISSING FEATURE THEORY

Johan de Veth, Bert Cranen & Louis Boves

A^{2}RT, Department of Language and Speech,
University of Nijmegen,
P.O. Box 9103, 6500 HD Nijmegen, THE NETHERLANDS

ABSTRACT

Acoustic backing-off was recently proposed as an operationalisation of missing feature theory for increased recognition robustness. Acoustic backing-off effectively removes the detrimental influence of outlier values from the local decisions in the Viterbi algorithm without any kind of explicit outlier detection. In the context of connected digit recognition over telephone lines, it is shown that with more than 30% of the static mel-frequency cepstral coefficients disturbed, acoustic backing-off is capable of reducing the word error rate by one order of magnitude. Furthermore, our results indicate that the effectiveness of acoustic backing-off is optimal when dispersion of distortions due to acoustic feature transformations is minimal.

1. INTRODUCTION

Recently, it was shown that missing feature theory can be used for improved robustness of automatic speech recognition (ASR) systems [1], [2]. According to missing feature theory, recognition performance in adverse conditions can be maintained at the level for undisturbed conditions provided that a sufficient number of acoustic features remain intact. The work in [2] provided a proof of concept: If an ASR device has prior information that some features are corrupted, and if its scoring procedure is such that corrupted features can be discarded, it is made very robust against distortions. In [3] we proved that missing feature theory can be used in conventional ASR without the need for prior knowledge about which features are corrupted. To this aim acoustic backing-off was introduced to limit the impact of possibly corrupted features in the scoring of the likelihood of alternative hypotheses. A similar robustifying effect was obtained as in [2]. Moreover, it was argued that the newly proposed acoustic backing-off (ACBOFF) method will work with any feature set, not just spectral features.

In this paper we extend the work of [3] in two directions. First, we present results of a study where we compared the effect of ACBOFF for disturbed and undisturbed feature vectors consisting of mel-frequency cepstral coefficients (MFCC's) as a function of the ACBOFF tuning parameter. Unlike the experiments described in [3], we applied the distortions to the acoustic features prior to channel normalisation and taking the first time-derivate. The results confirm the capacity of ACBOFF to restore recognition accuracy even when a substantial part of the MFCC's is disturbed.

Already in [3] we pointed out that the performance gain of ACBOFF is strongly dependent on the proportion of disturbed coefficients. Many physically realistic distortions in speech signals are local in the spectro-temporal space. However, many of the parameter transforms used in the front-ends of today's ASR systems imply some form of spectro-temporal smearing. As the second extension we therefore investigate how the performance gain of ACBOFF is related to two routinely applied parameter transformations, i.e., the discrete cosine transform (DCT) and channel normalisation (CN). The results of these experiments help to explain why these transforms may have undesirable side effects under some types of adverse acoustic conditions.

In section 2 we explain the theory underlying ACBOFF, section 3 describes the experimental set-up that we used, section 4 gives the major results and we formulate our conclusions in section 5.

2. THEORY

We assume that we have a set of independent measurements of a stochastic process at time instant t which constitute an observation vector $\mathbf{x}(t)$, with $\dim(\mathbf{x}) = K$. In addition, we assume that we have J distinct classes (states) $S_j, j = 1, \ldots, J$ from which the stochastic process originates.

Viterbi decoding needs some measure for *local distance* to identify the best path through the search space:

$$d_{loc}(S_j, \mathbf{x}(t)) = -\log[p(S_j)] + \sum_{k=1}^{K} \{-\log[p(x_k(t)|S_{jk})]\}, \quad (1)$$

where $d_{loc}(S_j, \mathbf{x}(t))$ is the local distance function (LDF), $p(S_j)$ is the probability of being in class S_j, and $p(x_k(t)|S_{jk})$ denotes the likelihood of observing feature value $x_k(t)$ according to coordinate k of class S_j.

Any procedure which limits the impact of distortions and other outliers on the LDF should help to diminish the effects of parameters with values that are widely beyond what was observed during training (cf. [2], [4], [5]). In a study in the field of speaker recognition [6] it was proposed to hard limit the cost function at $\mu \pm 3\sigma$. In [3] we proposed to limit the contribution of a –possibly corrupted– parameter observation to the LDF by means of a backing-off procedure. We compute the contribution $p(x_k(t)|S_{jk})$ in Eq. (1) as follows

$$-\log[p(x_k(t)|S_{jk}] \approx -\log[\alpha\hat{p}(x_k(t)|S_{jk}) + (1-\alpha)p_{0k}], \quad (2)$$

with α a backing-off value ($0 < \alpha < 1$) and $\hat{p}(x_k(t)|S_{jk})$ the parametric approximation of $p(x_k(t)|S_{jk})$. Mixtures of continuous probability density functions (pdf's) have appeared to be very powerful and effective in ASR devices to describe $\hat{p}(x_k(t)|S_{jk})$. p_{0k} is the (constant) probability that an arbitrary observation falls beyond the central portion of the distribution. The fact that we have chosen p_{0k} to be independent of state j, ensures that the contribution of a corrupted observation to almost all pdf's becomes equal and the parameter is effectively discarded for this frame. For a full explanation we refer the reader to [3]. The right hand side of Eq. (2) is a continuous and continuously differentiable function. This eliminates the need to branch towards qualitatively different processes if an observation exceeds some necessarily arbitrary threshold. Thus, we have effectively removed the need for explicit and error prone procedures for detecting outliers.

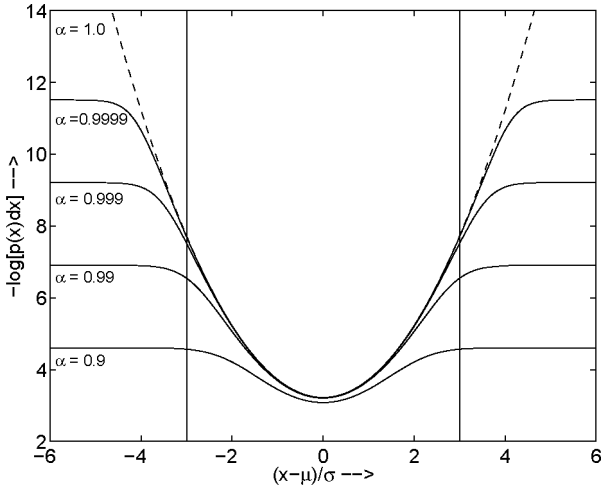

Figure 1: Contribution to local distance without (dashed line) and with (solid lines) ACBOFF for different values of α. Vertical lines indicate the boundaries of the region $|\frac{x-\mu}{\sigma}| \leq 3$.

Fig. 1 shows the effect on the LDF of ACBOFF for different values of α (solid curves), where we have used a single univariate Gaussian pdf for describing $\hat{p}(x_k(t)|S_{jk})$. As can be seen, decreasing the value of α results in decreasing the range of observation values where the contribution to the LDF $p(x_k(t)|S_{jk})$ is actually sensitive. In other words: By tuning α we tune the 'receptive field' of the emission pdf in our models.

3. EXPERIMENTAL SET-UP

We carried out our experiments with a connected digit recogniser trained for telephone speech. We artificially modified the acoustic vectors of the test utterances. The modifications we used are analytically tractable and easy to model, rather than physically realistic. In fact, these distortions are not intended to model specific real-life situations. Compared to adding certain types of noise to the original waveforms, our methodology allows us to better pursue our aims: (1) to investigate the potential power of our implementation of missing feature theory and (2) to investigate the impact of DCT and CN on the effectiveness of missing feature theory to robustify ASR systems.

3.1. Database

The speech material for our experiments was taken from the Dutch POLYPHONE corpus [7]. Speakers were recorded over the public switched telephone network in the Netherlands. For training we reserved a set of 480 connected digit strings, where each string contained six digits. For cross-validation during training [8] we used 240 utterances. The models were always evaluated with 671 independent test utterances.

3.2. Signal processing

A 25 ms Hamming window shifted with 10 ms steps and a pre-emphasis factor of .98 were used to calculate 24 filter band energy values, uniformly distributed on a mel-frequency scale (covering 0 - 2143.6 mel). Next, 12 MFCC's were computed. In addition we used the first time-derivatives (delta-MFCC's), log-energy (logE) and its first time-derivative (delta-logE), making for 26-dimensional feature vectors. Finally, we either applied cepstrum mean subtraction (CMS) or phase-corrected RASTA (pcR) [8] to the twelve MFCC's in order to normalise for channel variations. In both cases, we used the off-line version of the CN technique, i.e. using the whole utterance. The pre-processing steps in going from 24 log-energy bands to 12 channel normalised MFCC's and their corresponding delta's are shown in Fig. 2.

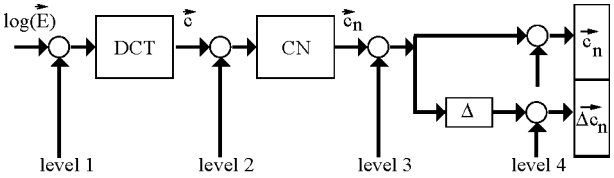

Figure 2: Schematic diagram showing the different acoustic pre-processing steps for coordinates depending on MFCC's in our feature vectors. The logE and delta-logE coordinates are not shown in this diagram. Δ indicates taking the first time-derivative. Level i indicates the position in the pre-processing where distortions may be introduced.

3.3. Acoustic distortion types

We used two different types of distortions for the experiments in this paper. As a first level 2 type distortion we randomly selected N_{sel} out of 12 MFCC coordinates. For all feature vectors in a test utterance the same N_{sel} coordinates were disturbed, but for each test utterance a new random selection was made. We denote this distortion type in short as RC. Using all available training data, we first determined the distribution of observation values for each individual coordinate. In each distribution, we determined a threshold value T_k such that 0.05% of the observations was lying above this threshold. A coordinate k selected to be distorted was assigned a value cT_k, with c a constant. We used $c = 0.75$ in all of the RC experiments in this paper and we always used $N_{sel} = 4$, which amounts to distorting more than 30% of the MFCC coordinates.
As the second type, we disturbed a sub-set of the log-energy band

values. This type of distortion takes place at level 1. For each band the maximum value was determined using all training data. For all frames in the test data the original value was replaced by the value corresponding to 10 dB below the maximum observed for that band if the original value was below this threshold, else the original value was kept. For the experiments in this paper we always applied the distortion to the first seven log-energy bands. This may be interpreted as a crude way of modeling a low-frequency additive noise (hence the short name: LF). Also in this case about 30% of coordinates was distorted.

3.4. Models

The ten words of the Dutch digits can be described with 18 context independent phone models. In addition we used four models for silence, very soft background noise, other background noise and out-of-vocabulary speech. For our most simple description, each phone unit was represented as a left-to-right hidden Markov model (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 next state. For these models the total number of different states was 66 (54 for the phones plus 12 for the noise models).

Contrary to the common modus operandi, we did not train multi-mixture Gaussian HMM's. Instead, right after the mixture split of our best single-Gaussian HMM's, we immediately rewrote each 3-state, double Gaussian HMM into the equivalent six-state, single Gaussian HMM with transitions allowed according to the topology-equivalence. In this manner we trained HMM's with a total of 66, 132 & 264 single Gaussian densities, with diagonal covariance matrices. For all experiments in this paper the models were trained only once, using undisturbed features.

4. RESULTS AND DISCUSSION

4.1. Level 2 distortions

In a first experiment we investigated the effectiveness of ACBOFF as a function of the backing-off parameter α. We compared recognition results for undisturbed and RC distorted feature vectors (cf. Fig. 3AB). As points of reference we also established recognition accuracy for undisturbed and RC distorted feature vectors without ACBOFF. The reference results are indicated in Fig. 3AB as the isolated short line segments in the right section of each pannel. In all cases, pcR was used for CN.

The results in Fig. 3AB clearly indicate that the RC distortion has a very large impact: WER jumps from 3.0% to 68.8% for HMM's with a total of 264 Gaussian pdf's. As can be seen, ACBOFF is highly effective in restoring much of the original recognition accuracy. For the 264 pdf HMM's ACBOFF reduces WER to 8.2%, which corresponds to a relative WER reduction of 88%. For this specific level 2 distortion the optimum value for α appears to be around 0.9 for each number of Gaussian densities that we tested ($-\log(1-\alpha) = 2.3$). At that optimum backing-off value, recognition performance for the undistorted feature vectors starts to deteriorate: The WER is increased to 3.7% for the 264 pdf HMM's. The gap between the two curves at this point may be interpreted as the effective loss in information due to the distortion. These results show that ACBOFF is a highly effective implementation of missing feature theory, capable of achieving a relative WER reduction well above 80%.

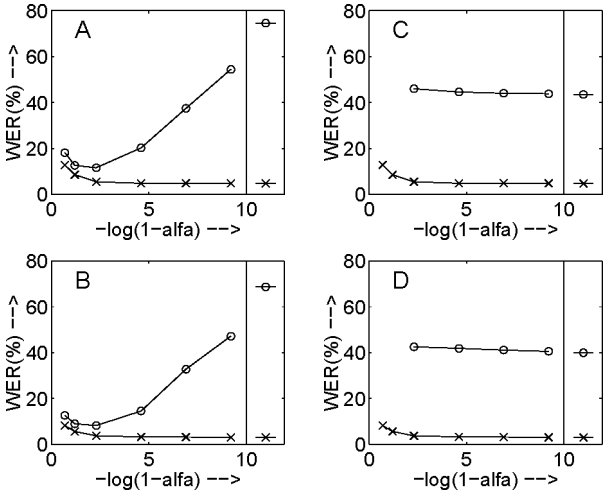

Figure 3: Recognition results as a function of backing-off parameter $-log(1-\alpha)$. Lines connecting '×': undisturbed condition. Lines connecting 'o': distorted condition. Results for A: HMM's with a total of 132 Gaussian pdf's, RC distortion; B: 264 pdf's, RC distortion; C: 132 pdf's, LF distortion; D: 264 pdf's, LF distortion.

4.2. Level 1 distortions

In the experiments reported on in [3] only level 4 distortions were applied. In the experiment described above, distortions were at level 2. None of these levels is physically realistic, but they have the advantage that they allow us to limit the distortions to a subset of the acoustic features used for recognition. In most real-world conditions, however, it is highly likely that distortions will be present in *all* cepstral coefficients: Even if only a small sub-set of the log-energy bands is distorted, the use of the DCT for obtaining the MFCC's will effectively smear the distortion over all cepstral coefficients. In addition to this within-vector dispersion, a temporal smearing will result from the use of CN and Δ. In the next two subsections we investigate the interaction between these spectral and temporal smearing phenomena and the effectiveness of ACBOFF as an implementation of missing feature theory.

Spectral Smearing To investigate the effects of spectral smearing we determined recognition performance with and without ACBOFF using the LF level 1 distortion and pcR for CN. The results are shown in Fig. 3CD. As can be seen the WER without ACBOFF for this level 1 distortion is well below the level found for the level 2 distortion shown in Fig. 3AB (compare 3C to 3A and 3D to 3B). For HMM's with 264 Gaussians in total we found WER=40.0%. Thus, judging from the WER values obtained without ACBOFF, it would appear that our LF level 1 distortion is less severe than the RC level 2 distortion that we investigated. As can be seen in Fig. 3CD, however, applying ACBOFF for our level 1 distortion does not give any improvement and even degrades the recognition performance somewhat. Of course, the level 1 and level 2 distortions that we applied are not related to each other in a way that

Table 1: WER results in % for different combinations of experimental set-ups.

features	$\alpha = 1$	$\alpha = 0.9$
original, pcR	3.0	3.7
original, CMS	3.0	4.2
disturbed, pcR	68.8	8.2
disturbed, CMS	90.1	7.4

would allow of straightforward analytical comparisons. Nevertheless, we are confident that the inability of ACBOFF (or probably any other implementation of missing feature theory, for that matter) to cope with this type of distortion can only be explained by the within-frame smearing caused by the DCT that affects all coefficients. This interpretation is in good agreement with [1], where it was already pointed out that missing feature theory does not allow application of an orthogonalisation transformation, because this violates the assumption that some of the feature values remain unaffected by the distortion. One possible way to reduce within-vector dispersion caused by DCT was recently suggested in [9].

Temporal Smearing As a final experiment we repeated experiment 1, but using CMS for CN instead of pcR. Using the RC level 2 type distortion in combination with CMS we again determined the optimum value for α and found $\alpha = 0.9$. For the models with 264 pdf's in total the WER results are shown for pcR and CMS together in Table 1 for the four different conditions that we evaluated. At a 95% confidence level the difference between pcR and CMS WER values when using optimal ACBOFF is not significant, for the undisturbed as well as the disturbed condition.

As can be seen in Table 1, ACBOFF is capable of reducing WER in the distorted condition with CMS by 92%. In other words: This is a reduction of WER by one order of magnitude.

When comparing the results for the two different CN techniques in Table 1, we observe a larger performance drop for CMS without ACBOFF. This may very well be explained by the fact that the CMS implementation that we used has an infinite impulse reponse. As a result, any local distortion will be distributed over the whole utterance. In the case of pcR, however, a local distortion will only be distributed over a limited number of neighbouring frames thanks to the finite impulse response of the pcR filter. As can be seen in Fig. 1, distortions causing larger outlier values make no difference when ACBOFF is being used, as long as the outlier values are lying outside the receptive field of the emission cost generator. This interpretation is well supported by the fact that WER is also restored in the case of CMS to a value that is not significantly different from the one obtained when pcR is used for CN. This result might very well be due to the fact that we used distortions such that a sub-set of the features was completely intact and all remaining feature-values were heavily distorted. Clearly, there are no doubt situations conceivable where some feature coordinates are mildly, others heavily and the rest not distorted. It remains an open question at this moment, if in such situations dispersion of distortions along the time axis is important. This will be one of the subjects of future research.

5. CONCLUSIONS

We studied the use of acoustic backing-off as a way to implement missing feature theory in the framework of an otherwise straightforward HMM recogniser. With ACBOFF the decoder does not need prior knowledge about which features are potentially distorted. In fact, it does not need any kind of explicit 'outlier detection'. When ACBOFF is applied to feature vectors suffering from inherently unpredictable distortions we could still improve the WER by more than one order of magnitude. In fact, we restored WER to the level expected for the original undisturbed feature vectors with some of the components removed. Our recognition experiments indicate that feature vector transformations in the front-end of the ASR should avoid dispersion of local distortions along the within-vector dimension as well as along the temporal dimension when using an implementation of missing feature theory for improved recognition robustness.

6. ACKNOWLEDGEMENT

This research 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).

7. REFERENCES

[1] M. Cooke, A. Morris & P. Green, "Recognising occluded speech". *Proc. ESCA Workshop on the Auditory Basis of Speech Perception*, Keele Univ., UK, pp. 297-300, 1996.

[2] R. Lippmann & B. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise". *Proc. Eurospeech-97*, pp. 37-40, 1997.

[3] J. de Veth, B. Cranen & L. Boves, "Acoustic backing-off in the local distance computation for robust automatic speech recognition". To appear in *Proc. ICSLP-98*, 1998. http://lands.let.kun.nl/literature/deveth.1998.2.html

[4] S. Dupont, H. Bourlard & C. Ris, "Robust speech recognition based on multi-stream features". *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 95-98, 1997.

[5] S. Tibrewala & H. Hermansky, "Sub-band based recognition of noisy speech". *Proc. ICASSP-97*, pp. 1255-1258, 1997.

[6] T. Matsui & S. Furui, "Comparison of test-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs". *Proc. ICASSP-92*, vol II, pp. 157-160, 1992.

[7] E.A. den Os, T.I. Boogaart, L. Boves & E. Klabbers, "The Dutch Polyphone corpus". *Proc. Eurospeech-95*, pp. 825-828, 1995.

[8] J. de Veth & L. Boves, "Channel normalization techniques for automatic speech recognition over the telephone". Accepted for publication in *Speech Communication*, 1998.

[9] S. Okawa, E. Bocchieri & A. Potamianos, "Multi-band speech recognition in noisy environments". *Proc. ICASSP-98*, pp. 641-644, 1998.