ACOUSTIC BACKING-OFF IN THE LOCAL DISTANCE COMPUTATION FOR ROBUST AUTOMATIC SPEECH RECOGNITION

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

In this paper we propose to introduce backing-off in the acoustic contributions of the local distance functions used during Viterbi decoding 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. It does so without the need for prior knowledge that specific features are missing. Acoustic backing-off avoids any kind of explicit outlier detection.
This paper provides a proof of concept of acoustic backing-off in the context of connected digit recognition over the telephone, using artificial distortions of the acoustic observations. It is shown that the word error rate can be maintained at the level of 2.5% obtained for undisturbed features, even in the case where a conventional local distance computation without backing-off leads to a word error rate $> 80.0\%$. The approach appears to be able to handle up to four independent corrupted features.

1. INTRODUCTION

Speech is a highly redundant communication medium. Most, if not all, information is coded in several different signal parameters. While this redundancy and cue-trading may make phonetic research more difficult, it certainly helps to make speech more robust under adverse acoustic conditions: If one or two features that can code a certain information item get drowned in background noise, chances are that enough other features are left that carry essentially the same information. To a large extent the lack of robustness of automatic speech recognition systems can be traced back to their rigid focus on a small subset of the possible parameters, which is only aggravated by the lack of effective techniques to discover that one or more of those focus parameters are affected by noise.
Recently, it was shown in [1] that the missing feature theory, derived from acoustic phonetics research, can be brought to bear on the issue of robustness of ASR systems. The work in [1] provided a proof of concept: If an ASR device has the 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. Of course, it is difficult to imagine that a real ASR system ever will have that prior information (beyond obvious information about channel bandwidth limits and coding). In this paper we want to take the proof of concept that missing feature theory can be used in 'conventional' ASR one step further. We show that prior knowledge about which features are corrupted is not necessary. Instead, statistical procedures that have proven their power in other problems in large vocabulary speech recognition can be used to limit the impact of possibly corrupted features in the scoring of the likelihood of alternative hypotheses, with the same robustifying effect as obtained in [1]. Moreover, since the method proposed in this paper is purely statistical in nature, it will work with any feature set, not just spectral features. In this paper we use MFCCs and their deltas, but any other feature set could have been used as well.
To be able to investigate the statistical properties of our approach we have analysed artificially induced corruptions of an initially clean parameter set. Moreover, we have used corruptions that are easy to model, rather than corruptions that are physically 'reasonable'. This allows us to investigate the effect of increasing the number of potentially corrupted parameters on the performance of the recogniser, and to make sure that the performance degrades gracefully.
In section 2 we explain the theory underlying our approach; section 3 describes the experimental set-up that was used to test the new approach, and section 4 gives the major results.

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 (e.g. [2]):

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

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 $x_k(t)$ belonging to coordinate k while in class S_j.
Mixtures of continuous probability density functions (pdfs) have appeared to be very powerful and effective in ASR devices to describe the likelihood $p(x_k(t)|S_{jk})$. So, as a parametric approxi-

mation of the likelihood $\hat{p}(x_k(t)|S_{jk})$, we may write

$$\hat{p}(x_k(t)|S_{jk}) = \sum_{m=1}^{M} c_{jkm} G(x_k(t), \mu_{jkm}, \sigma^2_{jkm}) \qquad (2)$$

with $\sum_m c_{jkm} = 1$, M the number of pdfs G in the continuous mixture and $\mu_{jkm}, \sigma^2_{jkm}$ denoting the m-th mean and variance describing coordinate k of class S_j, respectively. Since we assumed that we are working with independent scalar observations, we have variances here, not a covariance matrix. For clarity of the presentation, we will limit ourselves to a uni-variate, single Gaussian pdf. However, our approach generalises in a straightforward way to the case of multi-variate mixtures of Gaussians (or Laplacians or other parametric forms).

While the central portions of the parameter distributions in each state can be modelled accurately with almost any kind of mixture distributions, it remains questionable whether the same will ever hold for the tails of the distributions. By using the term 'distribution' we imply the assumption that all observations belong to some unique population, even is it is worthwhile to distinguish a number of sub-populations, each of which may be accounted for by one or more densities. However, if we take due account of the possibility of observations being distorted, and of the myriad number of ways in which distortions can affect parameter values, the question arises whether the borders of the distributions can at all be modelled accurately with the same set of mixtures that are trained to account for the central portion. Asking the question is paramount to answering it. Thus, we have looked for ways to model the borders of the distributions explicitly, and in a way independently from the central portions. The inspiration for the solution proposed in this paper came from the way the sparse data problem in language modelling is handled, where all observations beyond some distance from the centre of the distribution are given the same, finite count, instead of relying on the true count observed in the training corpus [3].

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

$$-\log[p(x_k(t)|S_{jk}] \approx -\log[\alpha\hat{p}(x_k(t)|S_{jk}) + (1-\alpha)p_{0k}], \quad (3)$$

with α a backing-off value. p_{0k} is the (constant) probability that an arbitrary observation falls beyond the central portion of the distribution. Unlike the approach in [6], Eq. (3) can still be interpreted as the $-\log$ of a true probability. This is a feature that we will need in future research.

Also, the right hand side of Eq. (3) 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.

Our model of acoustic observations is the sum of a mixture of (Gaussian) densities representing the central part of the distribution (i.e. the speech characterics as observed during training) and a uniform distribution that models the rest of the world. By doing so, we effectively admit that we have not enough training data to be able to discriminate between bad and worse outliers. It is better to treat them all in the same way, and attach constant penalty values to them, instead of accepting potentially extremely high penalties if a corrupted parameter happens to fall in the very far end of the tail of a (Gaussian) distribution which belongs to a different population than the observation at hand anyway.

Fig. 1 shows the effect of acoustic backing-off for $\alpha = 0.9999$ and $\alpha = 0.99$ respectively (solid curves). As points of reference, the boundaries of the region $|\frac{x-\mu}{\sigma}| \leq 3$ are indicated by vertical lines. Now, suppose that coordinate k of the current observation is an outlier; rather than giving rise to a very large penalty for the proper state, with our LDF the penalty will be bounded. Moreover, the same will hold for almost all other pdfs of almost all other states: Here too the acoustic contribution of $x_k(t)$ to the LDF reduces to $-\log[(1-\alpha)p_{k0}]$ (cf. Fig. 1). The fact that we have chosen p_{0k} to be independent of state j, ensures that the contribution of the corrupted parameter to almost all pdfs becomes equal and the parameter is effectively discarded for this frame. Of course, the corrupted parameter may happen to lay close to the central portion of the distribution in some completely unrelated states, giving rise to a small penalty. However, as long as most other parameters are uncorrupted, the total penalty for these states will still be relatively high, so that there is little risk that the corrupted parameter(s) will cause a strong preference for the wrong states. It is interesting and important that the effective elimination of a corrupted parameter will occur even if the corrupted value is well within the range of values for that parameter observed in all the training data. The only requirement that must be fulfilled is that it is shifted to the tail of most of the densities.

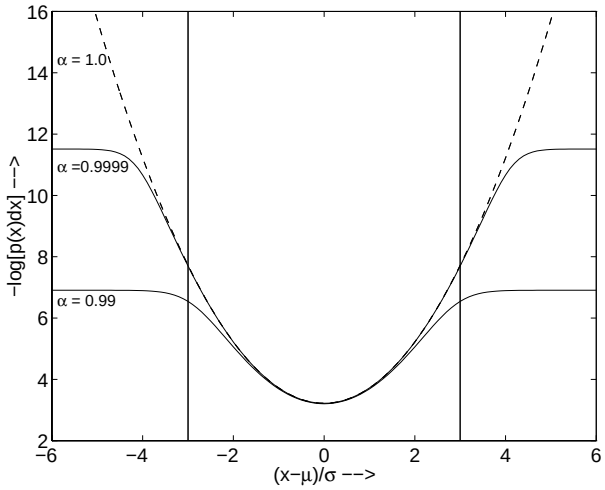

Figure 1: Contribution to local distance without (dashed line) and with (solid lines) acoustic backing-off for two different values of α (i.e. 0.9999 and 0.99). Vertical lines indicate the boundaries of the region $|\frac{x-\mu}{\sigma}| \leq 3$.

3. EXPERIMENTAL SET-UP

In order to test the effectiveness of a local distance function suggested in Eq. (3), we carried out an experiment with a connected digit recogniser trained for telephone speech. We artificially modified the acoustic vectors of the test utterances. We would like to stress that our artificially constructed distortions are not intended to model any real-life situation. Here we only want to prove that acoustic backing-off is capable to handle outlier observations in such a way that recognition performance degrades gracefully when the amount of distortion increases.

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. Handset and channel characteristics are not known; especially handset characteristics are known to vary widely. None of the utterances used for training or test had a high background noise level.
Among other things, the speakers were asked to read a connected digit string containing six digits (item "01" in the database). For training we reserved a set of 480 strings, i.e., 40 speakers (20 females and 20 males) from each of the 12 provinces in the Netherlands. An independent set of 911 utterances (461 females, 450 males) was set apart for testing. For cross-validation during training [8] we used a subset of 240 utterances taken from the test set (120 females, 120 males). For evaluation of the models we always used the 671 test set utterances that were not used during cross-validation.

3.2. Signal processing

Speech signals were recorded from a primary rate ISDN telephone connection and stored in A-law format. 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. The 24 triangular filters were uniformly distributed on a mel-frequency scale (covering 0 - 2143.6 mel) [9]. Finally, 12 mel-frequency cepstral coefficients (MFCCs) were computed. In addition to the twelve MFCCs we also used their first time-derivatives (delta-MFCCs), log-energy (logE) and its first time-derivative (delta-logE), making for 26-dimensional feature vectors. Finally, we applied cepstrum mean subtraction to the twelve MFCCs in order to normalise for channel variations.

3.3. Acoustic distortion types

We randomly selected 1 of the 26 coordinates of a feature vector. For that selected coordinate, the observation value was replaced by a new value which was calculated as follows. 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. The selected coordinate k was disturbed by assigning it the value cT_k, with c a constant. We always independently disturbed all feature vectors in a test utterance and did so for all test utterances. We used different values for c to be able to study the effect of scaling the distortion. In addition, for some of our experiments we disturbed not just one single coordinate, but did so for 2, 4, and 8 coordinates.

3.4. Models

The ten words of the Dutch digit set 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. Each HMM consisted of three states. The total number of different states was 66 (54 for the phones plus 12 for the noise models) for our most simple models. We trained HMMs with up to 528 Gaussian densities in total, each time with diagonal covariance matrices. The HMMs were strict left-to-right, with only self-loops and transitions to the next state. For all experiments reported in this paper, the models were trained only once using undisturbed features.

4. RESULTS AND DISCUSSION

In a first recognition experiment, the distortions were limited to one single coordinate randomly selected for each acoustic vector in our test utterances. The distortion scale factor c was varied in the range $0.1 - 1.0$. In this manner the distorted values are always still lying well within the range observed in the training data. It is reasonable to expect that the distorted values are in many cases outliers with respect to the Gaussian distribution to which the original value belonged. For comparison also undisturbed features were tested. We determined the word error rate (WER) for different HMM sets using LDF computation without and with backing-off. For the experiments with backing-off $\alpha = 0.9999$ was used. The results are shown in Fig. 2; the WER values at $c = 0.0$ show the results for the undisturbed features. Apparently, backing-off has virtually no effect on WER for undisturbed features (WER goes up from 2.4% to 2.5% for HMMs with 528 Gaussians).

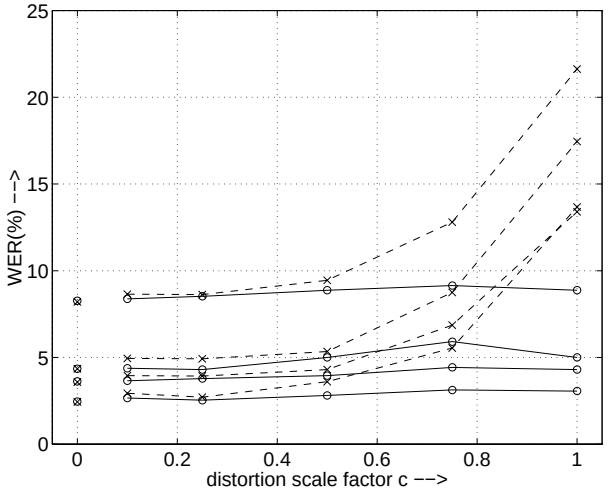

Figure 2: Recognition results for conventional LDF (dashed lines connecting '×') and LDF using backing-off with $\alpha = 0.9999$ (solid lines connecting '∘') for a single distorted coordinate. In both cases results are shown for HMMs with (from top to bottom) a total of 66, 132, 264 & 528 Gaussian densities.

The results in Fig. 2 clearly indicate that acoustic backing-off is very effective for this type of distortions. When backing-off is applied the WER remains at the level of the undistorted condition, whereas WER increases significantly without backing-off. Although WER differences are small when backing-off is applied,

it appears that recognition performance suffers most for $c = 0.75$.
In a second experiment we compared the recognition performance without and with backing-off for extreme distortions: Corrupted parameters were given the value $1.9T_k$. This corresponds roughly with the maximum value of the parameter ever observed in the training data. Without backing-off we found WER $= 88.8\%$ for the HMM set with 528 Gaussians. When we applied backing-off with $\alpha = 0.9999$ for the same model set we found WER $= 2.6\%$. Thus, acoustic backing-off is capable of maintaining the WER level at 2.5%, where LDF computation without backing-off leads to a WER value $> 80\%$.
Next, we determined recognition performance with and without backing-off as a function of the number of coordinates disturbed using $c = 0.75$ in all cases. The results shown in Fig. 3 are for HMM sets with a total of 528 Gaussian densities. Results for 66, 132, 264 Gaussians are not shown but are very similar. With $\alpha = 0.9999$ recognition starts breaking down at two distorted coordinates. With $\alpha = 0.99$, however, recognition performance is maintained at an acceptable level for distortions in up to four coordinates. Thus, the backing-off factor α should be diminished in order to longer maintain recognition performance at the level in the undistorted condition when more and more coordinates are disturbed.
Experiments are under way to determine the sensitivity of the recogniser performance to the value of α. Also, experiments are planned in which realistic distortions are applied to the speech signals prior to their transformation to MFCCs.

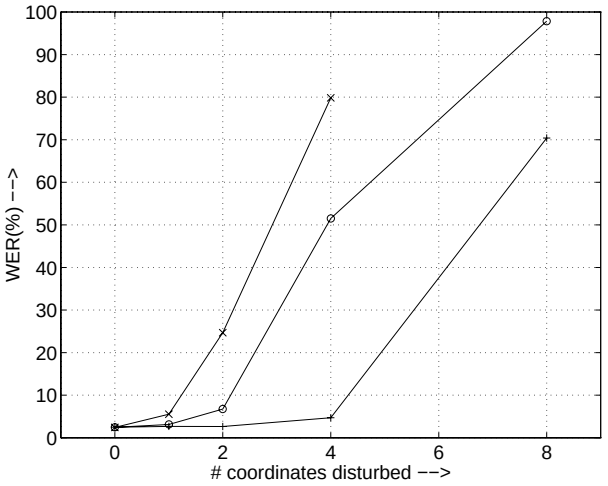

Figure 3: Effect of acoustic backing-off on WER as a function of the number of disturbed coordinates. Three different values of the backing-off factor were used: $\alpha = 1.0$ (i.e. no backing-off; indicated with $\times$), $\alpha = 0.9999$ (indicated with $\circ$) and $\alpha = 0.99$ (indicated with $+$). Results shown are for HMMs with a total of 528 Gaussian densities.

5. CONCLUSIONS

In this paper we proposed to use acoustic backing-off as a way to implement missing feature theory in the framework of an otherwise straightforward HMM recogniser. In our approach the decoder does not need prior knowledge about which features are potentially distorted. On the contrary, it does not need any kind of explicit 'outlier detection'. This property should make it suitable for handling real-world distortions due to background noise or bit errors in digital radio links.
If acoustic backing-off is applied to clean signals, the WER of the recogniser is not affected. In the simulation experiments reported in this paper the performance of the recogniser remained at the same level when increasingly large distortions were applied in up to four independent features. The performance of the original recogniser started to break down at moderate distortions in a single feature.
Our results show that the concept of missing feature theory to improve robustness of ASR is viable, and that it remains viable without the need for explicit decisions about which features are missing.

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).

6. REFERENCES

1. R. Lippmann & B. Carlson, 'Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise', in Proc. Eurospeech-97, pp. 37-40, 1997.
2. H.A. Bourlard & N. Morgan, 'Connectionist speech recognition, a hybrid approach', Kluwer Academic Publishers, Boston, 1994.
3. F. Jelinek, R.L. Mercer & S. Roukos, 'Principles of lexical language modeling for speech recognition', in *Advances in speech signal processing*, S. Furui & M.M. Sondhi eds., Marcel Dekker, New York, pp. 651-699, 1992.
4. S. Dupont, H. Bourlard & C. Ris, 'Robust speech recognition based on multi-stream features', in 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', in 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', in 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', in 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. Young, J. Jansen, J. Odell, D. Ollason & P. Woodland, 'The HTK book (for HTK Version 2.0)', Cambridge University, UK, 1995.