ACOUSTIC PRE-PROCESSING FOR OPTIMAL EFFECTIVITY OF MISSING FEATURE THEORY

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

A^{2}RT, Department of Language and Speech,
University of Nijmegen,
P.O. Box 9103, 6500 HD Nijmegen, THE NETHERLANDS
email: {J.deVeth, B.Cranen, F.deWet, L.Boves} @let.kun.nl

ABSTRACT

In this paper we investigate acoustic backing-off as an operationalization of Missing Feature Theory with the aim to increase recognition robustness. Acoustic backing-off effectively diminishes the detrimental influence of outlier values by using a new model of the probability density function of the feature values. The technique avoids the need for explicit outlier detection. Situations that are handled best by Missing Feature Theory are those where only part of the coefficients are disturbed and the rest of the vector is unaffected. Consequently, one may predict that acoustic feature representations that smear local spectro-temporal distortions over all feature vector elements are inherently less suitable for automatic speech recognition. Our experiments seem to confirm this prediction. Using additive band limited noise as a distortion and comparing four different types of feature representations, we found that the best recognition performance is obtained with recognizers that use acoustic backing-off and that operate on feature types that minimally smear the distortion.

1. INTRODUCTION

In automatic speech recognition (ASR), adverse acoustic conditions are likely to cause contamination of one or more components of the feature vectors that are computed in the front-end of the ASR-system. If no measures are taken to handle disturbed features differently from undisturbed features, it may be expected that recognition performance will drop. Recently, it was suggested that Missing Feature Theory (MFT) can be used to improve robustness of ASR under adverse acoustic conditions [1], [2], [3]. By disregarding unreliable acoustic features, recognition performance can almost be maintained at the level for undisturbed conditions, provided that most of the vital information is still present in the remaining undisturbed features.

In standard HMM recognizers, feature distributions are often modeled by means of Gaussian probability density functions. However, it is rather unlikely that the tails of a Gaussian distribution are reliable estimators of the less frequently occurring feature values. In [4], [5] it was proposed to model feature value observations by means of a mixture of two distributions: one obtained from the training data and another, uniform distribution which represents all feature values not seen during training. As it turns out, this idea is essentially the core rationale of statistical robustness theory [6]. In our proposal, referred to as *acoustic backing-off*, the local distance computation is implemented as the logarithm of a weighted sum of these two distributions; the weight assigned to either distribution can be varied so as to increase or decrease the contribution of the unseen values. We refer to the single parameter controlling the relative weight of the two distribution in the mixture as the acoustic backing-off parameter. In real-world testing conditions, it may be chosen a priori [4], [5].

Acoustic backing-off can be considered as an implementation of MFT which (1) is suited to be used in a conventional ASR system, (2) in principle allows one to use any feature representation as long as at least part of the acoustic feature vector is undisturbed, (3) contrary to the approach suggested in [2] does not require prior information about the corrupted features and (4) does not rely on an explicit detection mechanism for identifying disturbed feature vector elements as opposed to the approaches suggested in [7], [8].

In typical ASR systems, speech at the input is represented in some form of spectral representation. Most of the time, this *raw input feature* representation is not used directly to build statistical models, but various normalization and orthogonalization transforms are applied, e.g. gain normalization, channel normalization, Discrete Cosine Transform (DCT), Linear Discriminant Analysis (LDA). In this manner, features are obtained that are statistically more stable and/or allow for more efficient modeling. For clean speech data, these transforms generally improve recognition performance significantly. However, a complication arises when a subset of the raw input features are disturbed. In this case, the misleading information due to the disturbances which are present in a restricted number of raw features, will be smeared out over the entire normalized (orthogonalized) vector. If this happens, it may be expected that the effectiveness of MFT for fully recovering from the disturbances, is reduced. This is because the basic presupposition in MFT is violated: the disturbances must be such that only part of the feature vector is affected and the rest is still intact.

In an earlier experiment [5], we have shown this effect for a number of artificial disturbances. Acoustic backing-off appeared to be capable of restoring recognition performance when the MFCC features were disturbed directly. When a subset of the raw log energy features was disturbed by setting them to a fixed large value (i.e. before application of the DCT), however, the technique was no longer effective. The experiments in this paper are an extension of this preliminary experiment where the main goal was to study detailed characteristics of the method itself. Instead of artifically disturbing feature values, we now added band limited noise directly to the speech signal. There are two reasons for this. First, we wanted to make sure that acoustic backing-off is not only effective for the highly synthetic distortions that we used earlier, but that it is also effective against a somewhat more realistic disturbance, i.e. additive noise. Second, we wanted to investigate whether there is an interaction between different feature representations and the effectiveness of acoustic backing-off.

With the results obtained in this study, we intend to show that every possible effort should be taken to minimize the dispersion of disturbances. Although this holds true both for the within vector dimension and for the time (across vector) dimension. In order to keep the experiment sufficiently small given the four-page format of this contribution, this paper mainly focusses on the effects of within-vector smearing.

In sections 2 to 5, we describe our experimental set-up in more detail. In section 6 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 representations, with and without applying MFT in the form of acoustic backing-off. Finally, our conclusions are presented in section 7.

2. SPEECH MATERIAL

The speech material for our experiments was taken from the Dutch POLYPHONE corpus [9]. Speech was recorded over the public switched telephone network in the Netherlands. Among other things, the speakers were asked to read several connected digit strings. The number of digits in each string varied between 3 and 16. For training we reserved a set of 1997 strings (16582 digits). Care was taken so as 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. [10]) we used 504 digit string utterances (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 with regards to the number of males and females, the coverage of different regions in the country as well as to an equal number of tokens per digit. None of the utterances used for training or testing had a high background noise level.

3. ACOUSTIC FEATURES

In all our experiments we used mel-frequency log-energy coefficients (MFLECs) as the raw input feature representation. A 25 ms Hamming window shifted with 10 ms steps and a pre-emphasis factor of 0.98 were applied. 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 log-energy for each frame. These signal processing steps were performed using HTK2.1 [11].

From these raw input features we calculated four different feature representations, i.e.:

1. within-vector averaged mel-frequency log-energy coefficients (WVA-MFLECs)
2. mel-frequency cepstral coefficients (MFCCs)
3. sub-band mel-frequency cepstral coefficients (SB-MFCCs) [12], and
4. within vector filtered mel-frequency log-energy coefficients (WVF-MFLECs) [13].

As explained in more detail below, WVA-MFLECs and MFCCs are calculated from the entire vector of raw input features. As a consequence, any distortion in the raw input features is dispersed over all feature values that are used during recognition. SB-MFCCs and WVF-MFLECs are designed to avoid that distortions which are present in part of the raw input feature vector spread over the entire feature vector that results after preprocessing.

For the WVA-MFLECs, we computed the average within-vector log-energy value for each frame. This within-vector average (WVA) was subtracted from each of the original 16 MFLEC values yielding 16 WVA-MFLEC values. Additionally, we subtracted the average value (computed over the whole utterance) for all 16 WVA-MFLEC values as an implementation of a channel normalization (CN) technique. Finally, we computed the 16 corresponding time derivatives (delta-coefficients). Combining these with the 16 static WVA-MFLECs, log-energy and delta log-energy yielded 34-dimensional feature vectors.

In the case of MFCCs, (c1, ..., c12) were computed from the raw MFLECs using the DCT. Cepstrum mean subtraction (CMS) was then applied to the twelve MFCCs as a CN technique. We used the off-line version of this CN technique, i.e. the cepstrum mean was computed using the whole utterance. Finally, we computed the time derivatives and added these to the 12 channel normalized MFCCs. Together with log-energy and delta log-energy we obtained 26-dimensional acoustic feature vectors.

SB-MFCCs were computed by computing $(c_{1,1}, ..., c_{1,6})$ independently for the first 8 MFLEC values (covering 0 - 1218 Hz) and $(c_{2,1}, ..., c_{2,6})$ for the second 8 MFLECs (covering 1015 - 4000 Hz). Next, we proceeded analogously as with the MFCCs, 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 in this manner at 26-dimensional feature vectors.

WVF-MFLECs were computed by applying the filter $z - z^{-1}$ within each frame for coefficients $2 - 15$. Coefficients 1 and 16 were just copied. After this filter and copy operation, the mean value computed over the whole utterance was subtracted as a form of CN. Next the deltas were computed. The static and delta WVF-MFLECs were combined together with log-energy and delta log-energy to arrive at 34-dimensional feature vectors.

4. DISTORTIONS

In three independent experiments we added band limited, stationary noise to the speech signals so that SNR levels resulted of 20, 10 and 5 dBA respectively, i.e. both the speech and noise energy levels were weighted according to the A-scale [14]. 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 resulting raw input features would be contaminated by noise ($F_{low} = 375Hz$ and $F_{high} = 915Hz$). Furthermore, the value of the high cut-off frequency ensured that the noise distortions were limited to the first set of sub-bands used in the SB-MFCC feature representation.

5. HIDDEN MARKOV MODELING

The ten 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. 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 the next state. For these models the total number of different states was 63 (54 for the phones plus 9 for the noise models). We used HTK2.1 for training and testing HMMs [11]. We followed the cross-validation scheme described in [10] to determine the optimal number of Baum-Welch iterations. The more complex 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 such that connected digit strings, varying in length from 3 to 16 digits, could be recognised.

6. RESULTS AND DISCUSSION

In order to determine a proper reference system for each feature representation, we computed the word error rate (WER) in the undisturbed condition. WER $= (S + D + I)/N \times 100\%$, with N 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 this *clean speech* condition, we obtained WER values in the range of 2.4% (WVF-MFLECs) to 3.4% (WVA-MFLECs). We found no statistically significant differences between the WER values of WVA-MFLECs, MFCCs and SB-MFCCs. However, the WER values of the WVF-MFLECs representation were significantly lower than that of the other representations. This finding is in good agreement with observations reported in [13].

For each combination of SNR level and feature type, we evaluated the system performance using a recognition based on a conventional local distance function and a local distance function with acoustic backing-off. Based on earlier experience [5], we a priori fixed the value of the acoustic backing-off parameter in all experiments so that recognition performance in the clean condition did not suffer too much. The results for the four different feature representations are shown as a function of SNR in Fig. 1 for the conventional local distance function. The corresponding results for the local distance function with acoustic backing-off are shown in Fig. 2.

Focussing on the conditions where noise was added to the speech signals, two effects are clearly visible. First, recognition performance is better for the two feature representations that only partially smear distortions [i.e. SB-MFCCs and WVF-MFLECs (two rightmost bars)] compared to the two feature representations that smear distortions over all feature components [i.e. WVA-MFLECs and MFCCs (two leftmost bars)]. This observation holds both for the recognizer with the conventional local distance function (Fig. 1), and for the recognizer with a local distance function in which acoustic backing-off is applied (Fig. 2). Second, the recognizer with acoustic backing-off consistently yields better results in the noisy conditions compared to the recognizer with a conventional local distance function. This performance improvement is observed for all feature types that were tested.

The fact that representations which keep the disturbance limited to only part of the feature vector are better suited to base recognition on, is in good agreement with the results reported in [12], where SB-MFCCs were compared to MFCCs. Also the fact that acoustic backing-off is capable to increase recognition robustness in the presence of additive noise is in agreement with our expectations. If certain vector elements behave as outliers compared to the behavior that was observed during training, it is better to base recognition on those elements that were not affected. According to the set-up we propose, the feature vector components that don't fit the statistics of the training data are treated as misleading information and their effect is reduced during the computation of the best path in the Viterbi.

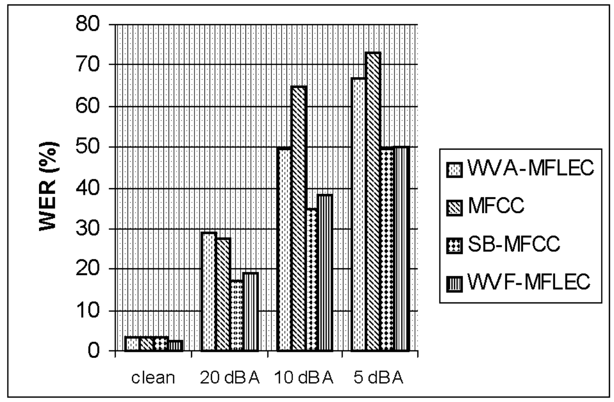

Figure 1: Recognition results as a function of SNR when using the conventional local distance function.

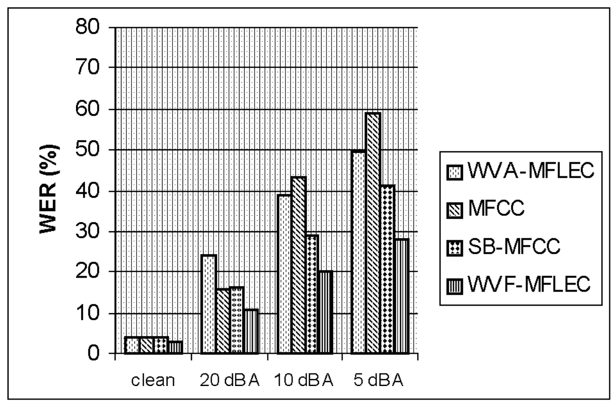

Figure 2: Recognition results as a function of SNR when using the local distance function with acoustic backing-off.

There appears to be no clear difference between the effectiveness of acoustic backing-off for representations which smear distortions over the entire vector and for representations that do not. At first glance, this finding seems to be in contradiction with results we found in earlier experiments [5] where we used artificial distortions. For instance, as mentioned before, one of the findings in those experiments was that acoustic backing-off was not effective at all with MFCC features when a subset of the MFLECs was set to a fixed, large value. However, that type of artificial distortion is so different from the distortions used in the current experiment, that comparison is not straightforward. With the artificial distortions, many MFCC coefficients were severely disturbed.

To illustrate how the band limited noise affects the MFCC coefficients in our current experimental set-up, we computed the normalized mean square error between the corresponding components of the disturbed and undisturbed MFCCs [15]. The result is shown in Fig. 3 for the condition with SNR = 20 dBA. As can be seen, the normalized mean square error is not evenly distributed over all cepstral coefficients. On the contrary, 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 the WER reduction observed when switching from a conventional local distance function to a local distance function with acoustic backing-off. The fact that the WER is not fully restored to the level observed in the clean condition can be understood by realizing that all coefficients are at least disturbed to some extent, i.e., not a single coefficient is still completely intact.

The observations above illustrate that an interaction exists between characteristics of the noise source and the way these affect the statistical properties of the features. It also illustrates that we lack a good method for predicting the effectivity of any method for robust ASR that can make use of detailed knowledge about how individual feature vector components actually used by the ASR-system are disturbed. Clearly, the normalized mean square error is not very well suited for this aim.

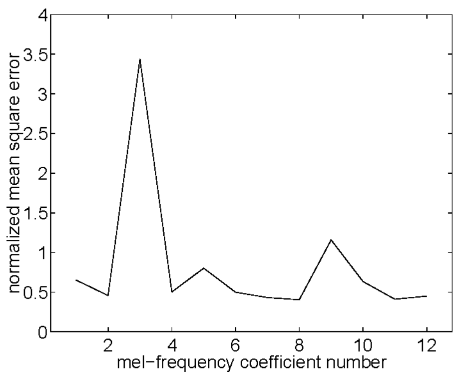

Figure 3: Normalized mean square error for the 12 MFCCs resulting from the SNR = 20 dBA distortion.

7. CONCLUSIONS

In our experiments, we used two representations that smear spectrally local distortions over all feature vector components and two representations that limit smearing to a sub-set of the feature vector components used for modeling and recognition. As a distortion we used band limited additive noise resulting in speech utterances with SNRs 20, 10 and 5 dBA. We found that the full smearing representations (WVA-MFLEC and MFCC) yield higher WERs than the representations that keep distortions limited to a subset of vector components (SB-MFCC and WVF-MFLEC). This is a clear indication that care must be taken in adverse conditions to choose a feature representation in which possible noise sources affect as few feature vector components as possible.

For all representations we investigated, acoustic backing-off appeared to be effective in improving noise robustness. We argued that this may be explained by the particular way in which the distorions are distributed over the different feature vector components. Some components are much more heavily distorted than others. Acoustic backing-off will limit the impact of the most severely affected outliers, so that recognition is effectively based on those features that are from a statistical point of view the least affected.

The two conclusions stated above are best reflected by the fact that WVF-MFLECs in combination with acoustic backing-off consistently gives the best results for all four SNR conditions that we studied. At $SNR = 5dBA$ the WER reduction is well above 40%.

Finally, we conclude that new methods need to be developed to assess the impact of mismatched training-test conditions on recognition performance. Due to the specific manner a given noise source may affect a certain feature vector, the insights gained from such (to the best of our knowledge as yet non-existent) mismatch assessment tools, are of key importance to further development of noise robust ASR.

ACKNOWLEDGEMENT

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

8. REFERENCES

1. M. Cooke, A. Morris & P. Green, 'Recognising occluded speech', in 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', in Proc. Eurospeech-97, pp. 37-40, 1997.
3. A. Morris, M. Cooke & P. Green, 'Some solutions to the missing feature problem in data classification, with applications to noise robust ASR', in Proc. ICASSP-98, pp. 737-740, 1998.
4. J. de Veth, B. Cranen & L. Boves, 'Acoustic backing-off in the local distance computation for robust automatic speech recognition', in Proc. ICSLP-98, pp. 1427-1430, 1998.
5. J. de Veth, B. Cranen & L. Boves, 'Acoustic backing-off as an implementation of missing feature theory', PP-TST report 81, 1999. http://lands.let.kun.nl/literature/deveth.1999.2.html
6. P.J. Huber, 'Robust statistics', John Wiley & Sons, 1981.
7. 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.
8. S. Tibrewala & H. Hermansky, 'Sub-band based recognition of noisy speech', in Proc. ICASSP-97, pp. 1255-1258, 1997.
9. E.A. den Os, T.I. Boogaart, L. Boves & E. Klabbers, 'The Dutch Polyphone corpus', in Proc. Eurospeech-95, pp. 825-828, 1995.
10. J. de Veth & L. Boves, 'Channel normalization techniques for automatic speech recognition over the telephone', Speech Communication, vol. 25, pp. 149-164, 1998.
11. S. Young, J. Jansen, J. Odell, D. Ollason & P. Woodland, 'The HTK book (for HTK Version 2.1)', Cambridge University, UK, 1995.
12. S. Okawa, E. Bocchieri & A. Potamianos, 'Multi-band speech recognition in noisy environments', in Proc. ICASSP-98, pp. 641-644, 1998.
13. C. Nadeu, J. Hernando & M. Gorricho, 'On the decorrelation of filter-bank energies in speech recognition', in Proc. Eurospeech-95, pp. 1381-1384, 1995.
14. J.R. Hassall & K. Zaveri, 'Acoustic noise measurements', Brüel & Kjær, Denmark, 1979.
15. J. Huerta & R. Stern, 'Speech recognition from GSM codec parameters', in CD-ROM Proc. ICSLP-98, no page numbers, 1998.