In: Proc. Workshop on Robust Methods for ASR in Adverse Conditions, Tampere, FINLAND, 1999, pp. 231-234.

MISSING FEATURE THEORY IN ASR: MAKE SURE YOU MISS THE RIGHT TYPE OF FEATURES

Johan de Veth, Febe de Wet, Bert Cranen & 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, F.deWet, B.Cranen, L.Boves} @let.kun.nl

ABSTRACT

In this paper we investigate acoustic backing-off as an operationalization of Missing Feature Theory to increase recognition robustness in adverse acoustic conditions. Acoustic backing-off effectively removes the detrimental influence of outlier values from the local decisions in the Viterbi algorithm. It does so without prior knowledge about the specific feature vector elements which are unreliable; thus, the technique avoids the need for explicit outlier detection. From the theory underlying Missing Feature Theory it appears that acoustic feature representations which smear local spectro-temporal distortions over all feature vector elements are inherently unsuitable. Our experiments in the context of connected digit recognition over the telephone are presented that confirm this prediction. Our results show that feature representations which minimize distortion smearing are most suited to be used in combination with Missing Feature Theory. Using additive band limited noise as a distortion, we found that acoustic backing-off can achieve a word error rate reduction of 44% when within vector filtered mel-frequency log-energy coefficients are used.

1. INTRODUCTION

In automatic speech recognition (ASR), adverse acoustic conditions are likely to cause contamination of one or more components of the incoming feature vectors. When a feature obtains unusual values (compared to training conditions) and if no measures are taken to handle these disturbed features differently from the 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 using only the reliable parts of the acoustic information and disregarding unreliable acoustic features, recognition performance can almost be maintained at the level for undisturbed conditions.

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. As a consequence, it might not be such a good idea to define the contribution to the local distance function used during the dynamic programming as a quadratic function over the entire feature value range. In [4], [5] it was proposed to model feature value observations by means of two distributions: the one obtained from the training data and a uniform distribution which represents all feature values not seen during training. Local distance computation interpolates between these two distributions; the weight assigned to either distribution can be varied so as to increase or decrease the contribution of the unseen values. This strategy was called *acoustic backing-off* and it was shown that it 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 [6], [7].

However, the application of MFT is not as straightforward as it might seem, since there appears to be an interaction between MFT, as applied during recognition, and the signal preprocessing steps associated with typical ASR systems [5]. Normalizing and orthogonalizing transforms are widely used in state-of-the-art ASR systems, e.g. gain normalization, channel normalization, Discrete Cosine Transform (DCT), Linear Discriminant Analysis (LDA). The main reason for using normalization transforms is that they yield statistically more stable feature values. As a result, speech can be represented more reliably and more efficiently, as reflected by improved recognition capability and faster training and recognition procedures, respectively. Orthogonalization is generally applied to remove correlations between raw spectral features so that a full-covariance matrix can be replaced by a more efficient diagonal variance matrix, i.e. its elements can be estimated reliably with less data. For clean speech data, these transforms generally improve recognition performance significantly. In this paper we will discuss why, under acoustically adverse conditions, simultaneous application of MFT on the one hand, and normalization and orthogonalization transforms on the other hand, may become undesirable. An intuitive understanding of the reasons behind this incompatibility may be obtained by considering the following reasoning.

The basic pre-supposition in MFT is that a feature vector can be decomposed into a part which is virtually unaffected and another part which contains distorted features. As long as the loss of information about the speech signal represented by the disturbed features is relatively small, MFT predicts that recognition performance can be maintained at a level which is comparable to the undisturbed case, simply by discarding the disturbed features. However, a complication arises when the raw incoming features are first transformed by means of an algorithm which uses all feature vector elements to calculate a transformed vector. 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, there is little hope that MFT can effectively help

in recovering from the disturbances.

In our opinion one of the challenges in building robust ASR algorithms is finding a proper combination of MFT and acoustic feature representations. The experiments in this paper 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, this paper mainly focusses on the effects of within-vector smearing.

In the rest of the paper, we will assume that the incoming speech is represented as a set of mel frequency log energy coefficents (MFLECs). To distinguish these input vectors from the feature vectors that result from pre-processing, i.e. those which are actually used for recognition, we will call these mel filter bank outputs *raw input features*. When talking about feature values we mean the vector elements that result *after* pre-processing.

The HMMs used during experimentation were based on four different feature representations, i.e.:

1. within-vector averaged mel-frequency log-energy coefficients (WVA-MFLECs)
2. mel-frequency cepstral coefficients (MFCCs)
3. within vector filtered mel-frequency log-energy coefficients (WVF-MFLECs) [8], and
4. sub-band mel-frequency cepstral coefficients (SB-MFCCs) [9].

Details about these feature representations will be given in section 3. For the moment it suffices to note that the first two of these representations (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. The last two representations (WVF-MFLECs and SB-MFCCs) are designed so that distortions which are present in part of the raw input feature vector do *not necessarily* spread over the entire feature vector that results after pre-processing. In other words, given the type of distortion applied, these representations guarantee that part of the feature vector remains unaffected.

The rest of this paper is organized as follows. First, in sections 2 to 5, we describe the experimental set-up that we used 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 representation techniques, 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 [10]. 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. [11]) 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

We used four different types of acoustic features for our experiments: within-vector averaged mel-frequency log-energy coefficients (WVA-MFLECs), mel-frequency cepstral coefficients (MFCCs), within-vector filtered mel-frequency log-energy coefficients (WVF-MFLECs) and sub-band mel-frequency cepstral coefficients (SB-MFCCs).

In each case we first computed acoustic feature vectors consisting of 16 mel-frequency log-energy coefficients (MFLECs) using the following set-up. Speech signals were recorded from a primary rate ISDN telephone connection and stored in A-law format. After conversion to the linear domain, 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 [12].

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. 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 exactly 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.

The 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

Ideally, what we are striving to find is an acoustic representation technique which is immune against broad band, non-stationary noise and not just band limited, stationary noise. However, we decided to start the investigation with a simplified problem in order to gain insight into the way the acoustic representations are affected by different kinds of noise.

We added band limited, stationary noise to the speech signals at a level of 5 dBA, i.e. both the speech and noise energy levels were weighted according to the A-scale [13]. 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 such that approximately one quarter of the resulting *raw input features* would be contaminated by noise. Furthermore, the value of the high cut-off frequency ensured that the noise distortions were limited to the first set of sub-bands in the case of 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 [12]. We followed the cross-validation scheme described in [11] 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 different recognition systems with 2, 4, 8 and 16 Gaussians per state (containing respectively 126, 252, 504 and 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) for the best HMMs according to the cross-validation development test set at 1, 2, 4, 8 and 16 Gaussians per state. The WER was defined as

$$WER = \frac{S + D + I}{N} \times 100\%, \qquad (1)$$

where N is the total number of words in the test set, S denotes the total number of substitution errors, D the total number of deletion errors and I the total number of insertion errors. At 16 Gaussians per state we obtained WER values in the range of 2.4% (WVF-MFLECs) to 3.4% (WVA-MFLECs). The reduction in WER in going from 8 to 16 Gaussians per state did not justify an additional mixture split at 16 Gaussians per state. The results we obtained at this working point are shown in Tables 1 to 4 for the four different feature sets that we studied (the figures in brackets indicate the 95% confidence intervals). As can be seen, the WER values of WVA-MFLECs, MFCCs and SB-MFCCs do not show substantial differences. However, the WVF-MFLECs representation yielded significantly better results. This finding is in good agreement with observations reported in [8] and can be explained by the characteristics of the WVF operation, i.e. (1) decorrelation and (2) variance equalization. For the recognition experiments reported below we always used HMM systems with 16 Gaussians per state.

Using a distortion of band limited noise at an overall SNR level of 5 dBA, we evaluated the system performance using a recognition based on a conventional local distance function. In addition, we evaluated the recognition performance for the clean and the disturbed condition, using a local distance function with acoustic backing-off. Based on earlier experience [5], we chose the value of the acoustic backing-off parameter (i.e. the parameter that controls to what extent the contribution to the local distance function is limited for extreme feature values) such that recognition performance in the clean condition did not suffer too much.

Table 1: WER results WVA-MFLECs.

	clean	$SNR = 5$ dBA
conventional	3.4 (0.4)	66.7 (1.0)
acoustic backing-off	4.0 (0.4)	60.7 (1.1)

Table 2: WER results MFCCs.

	clean	$SNR = 5$ dBA
conventional	3.2 (0.4)	73.8 (1.0)
acoustic backing-off	3.9 (0.4)	59.1 (1.1)

Table 3: WER results WVF-MFLECs.

	clean	$SNR = 5$ dBA
conventional	2.4 (0.3)	50.1 (1.1)
acoustic backing-off	2.7 (0.3)	28.0 (1.0)

Table 4: WER results SB-MFCCs.

	clean	$SNR = 5$ dBA
conventional	3.3 (0.4)	49.6 (1.1)
acoustic backing-off	4.1 (0.4)	41.2 (1.1)

Looking first at the results for the noisy condition using the conventional set-up, we note that the recognition performance suffers most for the two feature representations that smear spectrally local distortions over *all* feature vector components (see Tables 1 and 2): These WER values are in the vincinity of 70%. On the other hand, the two feature representations with only partially smeared distortions yield a very substantially lower WER of approximately 50% (see Tables 3 and 4). The improvement in WER in going from MFCCs to SB-MFCCs while using a conventional local distance computation is in good agreement with the observations reported in [9]. Thus, even whith a conventional local distance function without acoustic backing-off, limiting the dispersion of the distortions in the raw feature values to only a sub-set of the feature vector components helps to reduce the detrimental effect of the distortions. This is completely in keeping with the predictions of Missing Feature Theory.

Turning to the results where acoustic backing-off was applied, we notice first that recognition performance in the disturbed condition is significantly improved for all four feature representations at the cost of some loss in recognition performance in the

clean condition. Second, it can be seen that the best overall results are obtained for the two set-ups where acoustic backing-off is combined with a feature representation which only partially smears distortions: WER = 28.0% for WVF-MFLECs and WER = 41.2% for SB-MFCCs. This result shows that one can benefit most from an implementation of MFT when spectrally local distortions are kept local in the feature vector components of the representation used for modeling and recognition.

Finally, it can be observed that the WER reduction in the case of WVF-MFLECs is 44%, whereas in the case of SB-MFCCs it is 17%. Thus, acoustic backing-off appears to be more effective in the case of WVF-MFLECs. Most probably, this finding can be explained by the fact that the fraction of disturbed feature components within each feature vector is smallest in the case the WVF-MFLECs. For WVF-MFLECs the number of disturbed static WVF-MFLECs is 6, which is also the number of disturbed delta WVF-MFLECs. In addition, the log-energy and delta log-energy are disturbed. Thus, the distortions are present in 14 of the 34 feature vector components (corresponding to 41.2%). For the SB-MFCCs, the distortions are present in the first 6 sub-band cepstral coefficients, the corresponding deltas, in log-energy and in delta log-energy. In that case, 14 of 26 feature vector components are affected by the distortion (corresponding to 53.9%). This may prove to be an inherent advantage of WVF-MFLECs over SB-MFCCs. This hypothesis is presently under investigation for large vocabulary continuous speech recognition.

Of course, besides the relative amount of distorted feature vector components, the relative amount of information in the unaffected feature vector components must also be taken into consideration here. For instance, if it appears that the sub-band cepstra derived from the high frequency part of the MFLECs contain less information than the sub-band cepstra corresponding to the lower frequency half, the type of distortion applied in the experiments for this paper has biased the comparison. Nevertheless, it is tempting to speculate that the higher effectiveness of our implementation of MFT in the case of WVF-MFLECs compared to SB-MFCCs may be primarily attributed to the inherently smaller amount of smearing. Experiments are under way to investigate this issue further.

7. CONCLUSIONS

We investigated the effectiveness of acoustic backing-off as an implementation of MFT for four different acoustic feature representations when the speech utterances were distorted by band limited additive noise (SNR = 5 dBA). 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. For the two representations with full smearing we found that the effectiveness of acoustic backing-off as an implementation of MFT is limited. In both cases we found a WER at a level of 60%. For the two representations that only partially smear spectrally local distortions over all feature vector components, we found that recognition robustness is already significantly improved by using a conventional local distance computation. For both methods a WER at a level of 50% is found. Additionally, the WER is substantially improved when acoustic backing-off is applied. In the case of WVF-MFLECs, acoustic backing-off is capable of reducing the WER by 44% to an absolute level of 28.0%. We interpret our results as support in favour of the idea that limiting smearing of spectrally local distortions is a key factor in successful application of MFT.

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', submitted for publication.
6. 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.
7. S. Tibrewala & H. Hermansky, 'Sub-band based recognition of noisy speech', in Proc. ICASSP-97, pp. 1255-1258, 1997.
8. C. Nadeu, J. Hernando & M. Gorricho, 'On the decorrelation of filter-bank energies in speech recognition', in Proc. Eurospeech-95, pp. 1381-1384, 1995.
9. S. Okawa, E. Bocchieri & A. Potamianos, 'Multi-band speech recognition in noisy environments', in Proc. ICASSP-98, pp. 641-644, 1998.
10. E.A. den Os, T.I. Boogaart, L. Boves & E. Klabbers, 'The Dutch Polyphone corpus', in Proc. Eurospeech-95, pp. 825-828, 1995.
11. J. de Veth & L. Boves, 'Channel normalization techniques for automatic speech recognition over the telephone', Speech Communication, vol. 25, pp. 149-164, 1998.
12. S. Young, J. Jansen, J. Odell, D. Ollason & P. Woodland, 'The HTK book (for HTK Version 2.1)', Cambridge University, UK, 1995.
13. J.R. Hassall & K. Zaveri, 'Acoustic noise measurements', Brüel & Kjær, Denmark, 1979.