Analysis of Disturbed Acoustic Features in terms of Emission Cost

Laurens van de Werff, Johan de Veth, Bert Cranen & Louis Boves

A[2]RT, Dept. of Language & Speech, University of Nijmegen, The Netherlands

laurensw@sci.kun.nl, deveth@let.kun.nl, cranen@let.kun.nl, boves@let.kun.nl

Abstract

An analysis method was developed to study the impact of training-test mismatch due to the presence of additive noise. The contributions of individual observation vector components to the emission cost are determined in the matched and mismatched condition and histograms are computed for these contributions in each condition. Subsequently, a measure of mismatch is defined based on differences between the two histograms. By means of two illustrative experiments it is shown to what extent this emission cost mismatch measure can be used to identify the features that cause the most important mismatch and how in certain cases this type of information may be helpful to increase recognition accuracy by applying acoustic backing-off to selected features only. Some limitations of the approach are also discussed.

1. Introduction

One cause for training-test mismatch in Automatic Speech Recognition (ASR) is additive background noise that is present during recognition, but not during training. In principle, four strategies can be followed to alleviate the effect of such mismatch on recognition performance [1]: (1) use noise robust acoustic features, (2) use noise robust models, (3) use a noise robust scoring procedure for matching observations to models, or (4) use a combination of these.

Although an improvement of recognition accuracy in the mismatched condition proves the effectiveness of a noise robustness technique, it remains unclear whether the improvement was achieved by reducing the mismatch for all elements in the feature vector equally, or for a sub-set only. Also, in order to predict whether further improvements can be achieved and whether this robustness technique can be expected to work well in other noise conditions, one needs a diagnostic tool. The tool should allow to investigate the relative contribution of each feature vector component to the overall training-test mismatch. Such a tool would be particularly important for guiding our *acoustic backing-off* work [2,3]. Acoustic backing-off is a robust scoring procedure that aims to reduce the impact of distorted feature values on the decoding process. It focuses on those feature values whose properties clearly deviate from the distribution of observations seen in the clean condition, which give rise to high contributions to the emission cost. In order to better understand where differences in recognition performance stem from and to determine to what extent a beneficial effect may be expected from acoustic backing-off, it is important to be able to determine the relative proportions of outlier values for each feature and to relate these to recognition accuracies obtained in clean and disturbed conditions, with and without applying backing-off. We therefore decided to develop an analysis tool with which the training-test mismatch conditions can be visualized per feature vector component in terms of emission cost (EC) distributions. In this paper we discuss the properties of one such tool in the context of our research on robust scoring techniques.

The rest of the paper is organised as follows. Section 2 describes the analysis tool. In Section 3 the experimental set-up of several different training-test mismatch conditions is described. Section 4 discusses the results of analysing the EC contributions of these mismatched conditions. Finally, the main conclusions are given in Section 5.

2. Emission Cost Contributions

We want to study how a particular noise distortion affects the scoring during recognition, by investigating how the contributions of each feature vector component to the total score are distributed. The emission cost per frame can be easily obtained along the optimal path once the dynamic programming has finished: When the acoustic models in the ASR system are described as mixtures of continuous density Gaussian probability density functions (pdfs), and the vector components are assumed to be statistically independent so that diagonal covariance matrices can be used, the emission likelihood for each {acoustic observation vector, state} pair along the best Viterbi path becomes

$$p(\vec{x}(t) \mid S_j) = \sum_{m=1}^{M} w_{jm} \prod_{k=1}^{K} G_{jmk}(x_k(t)) \tag{1}$$

where $\vec{x}(t)$ denotes the acoustic vector, S_j is the state considered, w_{jm} denotes the m^{th} mixture weight for state S_j, and G_{jmk} the k^{th} component of the m^{th} Gaussian for state S_j.

With a mixture of M Gaussians contributing to the emission likelihood it is impossible to write the emission cost [i.e., $-\log(\ p(\vec{x}(t) \mid S_j))$] as a sum of K independent contributions of individual feature vector components. However, replacing the sum of mixture components by the maximum over the weighted mixture components has a negligible effect on recognition performance [4]. Replacing the sum operator in Eq. (1) by a maximum operator, makes it possible to write the EC as a sum of contributions of individual feature components:

$$EC = -\log(w_{jb}) - \sum_{k=1}^{K} \log[G_{jbk}(x_k(t))] \tag{2}$$

with w_{jb} the weight and G_{jbk} the k^{th} component of Gaussian b ('best') in state S_j that was found to be most likely in the weighted mixture of Gaussian components.

Histograms can be made of the contributions of individual features in Eq. (2) (i.e., the $-\log(G_{jbk})$ terms) both for the clean and the mismatched condition. Denoting the EC distributions for the k^{th} feature component in the clean condition as $H_{k,clean}$

and for the mismatched condition as $H_{k,mis}$, a measure of mismatch can then be defined as follows:

$$M_k = \sum_{n=1}^{N} a^n \left| H_{k,mis}(n) - H_{k,clean}(n) \right| \quad (3)$$

where N is the number of bins in the histogram of EC contributions, and a denotes a weighting factor allowing to emphasize large EC contributions more than small ones[1]. Differences between the two histograms for small and large values of the EC are equally important when $a = 1$. However, since in a robust scoring technique like acoustic backing-off differences between the two histograms for large EC contributions are potentially more interesting (a rather arbitrarily chosen) $a = 1.1$ was used throughout this paper.

3. Experimental set-up

To evaluate the usefulness of the measure defined in (3), two experiments were carried out. In the first experiment training–test mismatch was created by artificially corrupting data from the Polyphone database. This experiment mainly serves as a sanity check in a well-controlled environment, and to illustrate the main properties of the measure. The purpose of the second experiment was to explore the ability of the mismatch measure to select the most suitable candidate feature(s) for acoustic backing-off, if any, for more 'realistic' distortions. A small subset of the Aurora2 database [6] was used.

3.1. Experiment #1: Artificially distorted Polyphone data

In the Dutch Polyphone corpus, speech was recorded over the public switched telephone network in the Netherlands [5]. Among other things, speakers were asked to read a connected digit string containing six digits. 480 strings were used for training and 671 different strings for testing. Both training and test set were balanced with regards to the number of males and females, and regions in the country. None of the utterances used had a high background noise level.

The acoustic vectors consisted of 14 filter-bank log-energy values, computed from 25 ms Hamming windows with 10 ms steps and a pre-emphasis factor of 0.98. The 14 triangular filters were uniformly distributed on the Mel scale (covering 0 – 2143.6 Mel). For the 14 log-energy coefficients the average log-energy (computed over a whole utterance) was subtracted as channel normalization. Δ-coefficients were added, yielding 28-dimensional feature vectors.

For modelling the digits, 18 context-independent phone-based hidden Markov models (HMMs) were used, consisting of 3 states, with only self-loops and transitions to the next state. Each state was modelled as a mixture of four Gaussian probability density functions, with diagonal covariance.

To create a well-defined mismatch condition an artificial distortion was introduced to log-energy bands 6, 7, and 8 (centres at 799, 1002, and 1232 Hz). The original value was replaced by the value corresponding to 4.3 dB below the maximum observed for that band if the original value was below this threshold, else the original value was kept. The replacement was done for each observation vector in the test data. This type of distortion may be interpreted as a crude way of modelling band-limited, additive noise. In the worst cases, just over 21% of the feature values were affected.

3.2. Experiment #2: Noisy Aurora2 data

The Aurora2 database is a noisified version of the TI-digits database [6]. In this experiment we only used the clean condition and the condition with subway noise added at an overall SNR of 10 dB.

Before the calculation of the actual acoustic vectors, the noise reduction scheme described in [7] was applied. Next, 12 Mel-cepstrum coefficients (c_1, …, c_{12}) were computed with the standard front-end that comes with Aurora2, together with overall log-energy [6]. Next, cepstrum mean subtraction was applied to c_1, …, c_{12} using the full length of each recording. Finally, Δ-coefficients and ΔΔ-coefficients were computed (both over a window of 9 frames), for an eventual number of 39 acoustic features.

For the Aurora2 digits a standardised ASR system is supplied, which consists of strictly left-to-right whole word models [6]. Each model consists of 16 HMM states, modelled as a mixture of three Gaussian probability density functions with diagonal covariance matrices.

4. Results and discussion

4.1. Artificial distortions using Polyphone

The artificial distortions for log-energy bands 6 - 8 affect the distributions of these parameter values quite severely. Since all values smaller than 4.3 dB below the maximum were replaced by a fixed value, the histograms of the distorted log-energy coefficients have a very narrow peak. This distribution deviates so much (both in terms of mean and standard deviation) from the clean parameter value distribution that during the decoding process, one should expect to see appreciable EC-contributions for most of the frames. Consequently, we expect that the mismatch measure defined in Eq. (3) is able to diagnose bands 6 - 8 as the main cause of the mismatch.

For the corresponding Δ-coefficients the distortion is probably less harmful. Since the majority of the features 6 - 8 contain a fixed value their corresponding Δ-coefficient distributions show an extremely high amount of zeros. Because the original clean distributions also have a mean value close to zero and a small standard deviation, not many large EC contributions are expected for the Δ-coefficients.

To test these hypotheses, all test utterances were recognized with an HMM system using the maximum mixture component approximation[2]. For the clean and the distorted condition, the EC contributions according to the best mixture component along the optimal path were pooled over all test utterances. Four illustrative examples of the resulting EC distributions are shown in Fig. 1. The left column depicts the EC-distributions without (dashed) and with (solid) distortion for an undistorted coefficient (log-energy of band #4) and its corresponding Δ-coefficient; the right column depicts the EC-distribution of a distorted coefficient (log-energy of band #7) and its corresponding Δ-coefficient.

Clear differences between all EC distributions of the clean and disturbed condition can be observed, even for the two unaffected components (panels A and C). The fact that the EC-contributions can differ for undisturbed features may be explained by the fact that the optimal Viterbi path for an utterance in the disturbed condition can differ from the path in the

[1] Because we were mainly interested in the relative EC contributions of individual vector components, the contribution of the -log(w_{ib}) term in Eq. (2) was disregarded in the subsequent analyses.

[2] When switching from the sum to the max approximation the word accuracy did not change: 88.6% (clean data), 56.0% (disturbed data).

clean condition. As a result, the EC histogram in the distorted condition may contain contributions from HMM states that were not part of the optimal path in the clean condition. Turning to EC-distributions of the distorted feature components, it can be seen that there are large differences between the histograms of EC contributions for the static component. Fig. 1B shows that there are many more frames in the test set where feature #7 has an EC contribution of approximately 5 than in the training set. As expected, the differences are much smaller for the corresponding Δ-component (Fig. 1D).

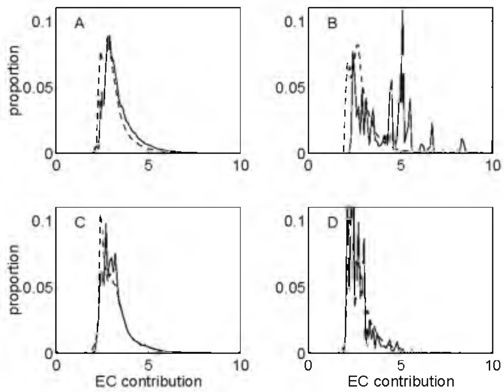

Figure 1: Distribution of EC contributions before (dashed line) and after (solid line) application of the distortion. A. Log-energy coefficient in band #4. B. Log-energy in band #7. C. Δ-log-energy of band #4. D. Δ-log-energy of band #7.

To estimate the impact of the distortion for individual feature vector components, the mismatch defined in Eq. (3) was determined. The result is shown in Fig. 2A. As expected on the basis of the differences in the histograms, the static components that were distorted (6 - 8) have a large mismatch value (the value for log-energy band 6 is clipped, reaching an actual value of 779). Somewhat surprisingly, also a relatively large mismatch was found for coefficients 5 and 9. Most probably, this is due to the fact that we used log-energy coefficients, which are known to show a high degree of co-variation, especially between neighbouring features.

To test whether the mismatch measure (3) can provide extra insights in the effectiveness of robust distance computation, two extra recognition experiments were run with acoustic backing-off. In the first one, the robust local distance function was applied to all coefficients; in the second one, acoustic backing-off was restricted to coefficients 6 - 8. Table 1 shows the corresponding recognition accuracies for the clean and distorted test utterances. As can be observed, application of acoustic backing-off for all coefficients does help to improve recognition performance in the mismatched condition. However, selective application of acoustic backing-off to only those three coefficients for which the EC mismatch was highest allows to fully restore the accuracy to the level observed for the clean condition.

Table 1: Recognition accuracies for clean and mismatched test utterances using different distance computation set-ups.

distance computation	*Clean*	*mismatched*
Conventional	88.6 %	56.0 %
Robust, all coefficients	88.3 %	80.5 %
Robust, only coefs 6,7,8	88.4 %	88.4 %

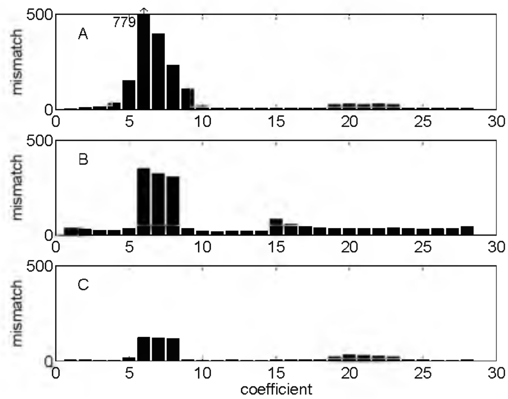

Figure 2: Mismatch as a function of the feature vector component. Components 1 - 14 correspond to log-energy band coefficients; components 15 - 28 correspond to the Δ-log-energy coefficients. A: Conventional distance computation. B: Robust distance computation for all coefficients. C. Robust distance computation restricted to log-energy coefficients 6 - 8.

Fig 2B shows that applying backing-off to all coefficients indeed reduces the EC mismatch for coefficients 5 - 7 and for coefficient 9, but at the same time increases the mismatch for all other coefficients. Thus, using a robust distance computation for all coefficients introduces an unintended EC mismatch for coefficients that are not disturbed (as well as for coefficient 8 in this particular case). Due to this newly introduced mismatch, only a limited gain in recognition accuracy is obtained. Fig 2C shows that if acoustic backing-off is only applied to the features that were most affected according to the mismatch measure, the unexpected EC mismatch increase is absent and most of the EC mismatch values are reduced.

4.2. Mismatched conditions in Aurora 2

The findings in section 4.1 suggests that an EC mismatch characterization as shown in Fig. 2 could be potentially helpful for selecting the feature components where robust EC computation would be most effective. To test this hypothesis, a similar procedure was applied for the Aurora2 database. For this data the EC mismatch appeared to be particularly large for overall log-energy (EC mismatch = 346) and for c_4 (EC mismatch = 163), while the remaining 37 feature components showed relatively small EC mismatch (values < 64). Based on these findings, three recognition experiments were run with robust distance computation: (1) for all coefficients, (2) for log-energy only, and, (3) for c_4 only. The results are shown in Table 2. The accuracy for the clean condition was 99.2 %.

Table 2: Recognition results for Aurora2 with 10 dB SNR subway noise; different distance computation set-ups.

Distance computation	*Del*	*Subst*	*Ins*	*Accuracy*
Conventional	70	199	26	90.9 %
Robust, all coefs	100	187	26	90.4%
Robust, only log-E	270	162	12	86.4 %
Robust, only c_4	58	187	28	91.6 %

As can be seen in Table 2, recognition performance (column 5) deteriorates both when the robust local distance function is applied to all coefficients and to log-energy only. Thus, in contrast to experiment #1, application of acoustic backing-off to the component showing the largest mismatch does not improve recognition performance. On the other hand, applying acoustic backing-off to the coefficient that gave rise to the second largest mismatch (c_4) does improve accuracy.

The fact that using acoustic backing-off during the scoring of the log-energy parameter does not improve recognition performance may seem strange at first glance, the more so because a detailed inspection of the mismatch revealed a decrease for almost all other vector components (not shown) in the relative EC contribution; for those few that did increase, the increase was small. Apparently, a decreased mismatch along the optimal Viterbi path (as defined in Eq. (3)) does not guarantee fewer recognition errors.

On second thought, this observation can be understood. The main reason for introducing acoustic backing-off was to deal with situations where incidentally a vector component has a value that is very unlikely according to the trained distributions, but where it is very difficult to estimate how unlikely. Especially if one suspects training-test mismatch the observed value might be distorted due to an unknown process of which one is unable to gather proper statistics. In these cases one should avoid the decoding process to be guided by the distance between observation value and distribution means of competing states [2]. The current situation is different, though. Inspection of the errors (Table 2: columns 2 - 4), shows that the number of substitution and insertion errors decreases, but that this effect is counteracted by a larger increase in the number of deletion errors. Due to the use of a robust distance measure, a speech observation vector is more often mistaken for silence: reducing the cost associated with a large log-energy mismatch, and, maybe worse, fixing this cost at a certain level, has caused an increase in the confusability between speech and non-speech sounds. Observation vectors that were classified as 'definitely *not* silence' without acoustic backing-off (using information from log-energy) can now unjustly be assigned to silence, because mainly spectral similarity is taken into account.

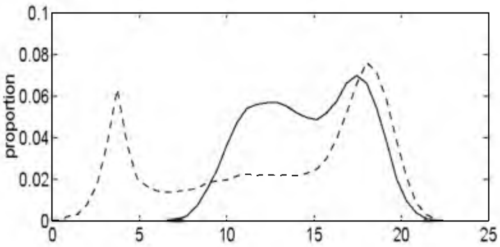

Figure 3: Distribution of the log-energy values in the clean (dashed line) and noisy condition (solid line).

The distribution of the log-energy in the clean and noisy condition, shown in Fig. 3, may help to understand this. In the clean condition the distribution of log-energy in silence will account for the left mode of the dashed curve. As is clear from the solid line, there are no noisy speech sounds that look similar to silence in the clean situation (explaining why we have high EC contributions for this parameter). Despite the mismatch, however, log-energy could still be used to sort sounds in low-intensity and high-intensity sounds. By applying acoustic backing-off this distinction is discarded.

5. Conclusions

A tool for analysing training-test mismatch was proposed, based on histograms of the EC contributions (determined along the Viterbi alignment path) of individual feature vector components over an entire test set. The mismatch is derived from the difference between the histograms of EC contributions for the clean and the mismatched condition. It was found that the mismatch measure is capable of determining whether a given noise type causes some feature components to contribute more to the overall mismatch than others. One way to use this information is to identify candidate features for selective acoustic backing-off. More extensive research with other databases and noise types is needed to show whether any regularity can be discovered in the features that show the largest mismatch.

Recognition experiments in which the impact of the selected components was reduced by means of our robust distance scoring technique, showed that the proposed selection method indeed *can* improve performance (e.g. in the artificially distorted Polyphone coefficients and the c_4 component in the Aurora2 case). However, this does *not always* constitute an effective way to improve performance. In particular, when applying acoustic backing-off to log-energy in the Aurora2 data, we found that the number of confusions between silence and speech sounds was increased. We believe that it was unwise to apply acoustic backing-off for the log-energy Gaussians that model silence. In order to test this hypothesis the diagnostic tool needs to be extended so that it can visualize EC contributions for specific models, states or mixture components.

More generally, our results suggest that acoustic backing-off should not be applied when outliers are systematically tied to specific states or models. In order to apply robust statistics they must be incidental. If they are not, other measures (like model adaptation) are in order. Extending the diagnostic tool so that EC-contributions can be visualized as a function of time, mixture component, state, or model will help to detect such situations.

Acknowledgement

This work was partially supported by the SMADA European project. The SMADA project is partially funded by the European Commission, under the Action Line Human Language Technology in the 5th Framework IST Programme.

References

[1] de Veth, J., Cranen, B., and Boves, L., "Acoustic features and distance measure to reduce vulnerability of ASR performance due to the presence of a communication channel and/or background noise", in: *Robustness in language and speech technology*, J.-C. Junqua and G. van Noord (Eds), Kluwer, 9 – 45, 2001.

[2] de Veth, J., Cranen, B., Boves, L., "Acoustic backing-off as an implementation of Missing Feature Theory", *Speech Comm.* 34, 247 – 265, 2001.

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

[4] De Mori, R. *Spoken Dialogues with Computers*, Academic Press, London, 1998.

[5] den Os, E., Boogaart, T., Boves, L., Klabbers, E. "The Dutch Polyphone corpus", *in Proc. Eurospeech-1995*, pp. 825-828, 1995.

[6] Hirsch, G., and Pearce, D., "Second experimental framework for the performance evaluation of speech recognition front-ends", *STQ Aurora DSR Working Group*, document AU/231/00, 2000.

[7] Noe, B., Sienel, J., Jouvet, D., Mauuary, L., Boves, L., de Veth, J. and de Wet, F. "Noise reduction for noise robust feature extraction for distributed speech recognition", *in Proc. Eurospeech-2001.*