Histogram Normalisation and the Recognition of Names and Ontology Words in the MUMIS Project

Eric Sanders, Febe de Wet

Radboud University Nijmegen, the Netherlands
[E.Sanders,F.deWet]@let.kun.nl

Abstract

The automatic transcription of German football commentaries and the analysis thereof are described. Histogram normalisation was used to improve the transcription of the very noisy data. The recognition of player names and ontology words was also investigated, since these are of crucial importance for the information retrieval task for which the transcriptions were used.

1. Introduction

The aim of this study was to improve the automatic transcription of spoken football commentaries using a pre-existing continuous speech recogniser by means of histogram normalisation (HN) and to investigate how well the most important elements in the commentaries, i.e. names and words from an ontology, are recognised. Part of the research was carried out in the framework of the FP5 IST project MUMIS (MUlti-Media Indexing and Searching environment) (2000-2002) [6]. In this project, a demonstrator was built that can retrieve specific video and sound fragments from recordings of football matches of the EURO-2000 championship, based on several sources of information, such as newspaper reports, tickers, subtitles, internet, teletext, and the spoken commentaries of the television broadcast of the matches. A merging tool was developed to combine information from the different sources [3]. The "formal annotations" that are used by the merging tool consist of words from an ontology of the football domain, created within the project [8], and the names of the players (as well as trainers and referees). We focus on recognition performance of these words and names in this paper, because they are crucially important for the retrieval task.

The television broadcast commentaries take a special place amongst all the information sources that are used in the project, because they provide the most detailed and accurate mapping between the transcription and the moment that an event occurred [8]. However, since they are also the only source of oral information, they need to be transcribed before they can be used by the merger. In [12] and [10] experiments are described that were aimed at optimising the automatic transcription of the commentaries by means of a speech recogniser. The spoken commentaries were extremely noisy, since the noise from the audience was mixed with the recordings of the commentators' voices. The high level of background noise made it extremely difficult to create accurate annotations of the data automatically. The word error rates (WER) obtained on the data typically varied between 50 and 80 %.

In this paper we report on attempts to improve the overall recognition rate by using a technique that has achieved promising results in previous studies on robust automatic speech recognition (ASR), i.e. histogram normalisation (HN) [13]. In the next section, the design of the experiments is presented. In the third section, the histogram normalisation technique we used and the results are described. In the fourth section the recognition performance of names of players and ontology words are compared to the overall recognition performance. In the fifth section we describe a few interesting points we came across during the error analysis of our results. In the final section we draw our conclusions.

2. Experimental setup

2.1. Material

The MUMIS data comprise commentaries of six matches in Dutch, three in English and 21 in German. For the experiments described here, only the commentaries in German are used, because for this language the largest amount of data was available. Three of the 21 matches were discarded, because the speech signals contained too many distortions. Not all matches had the same commentator, but within a match, most of the commentary was by one person. In case the recordings were in stereo, the channel with the best sound quality was used. The recordings of 13 matches were provided by the German TV company that broadcasted the events, while the remaining five matches were recorded on a home video system (either VHS or DV).

All speech data was manually transcribed following the annotation rules of the spoken Dutch corpus (CGN) [1]. The speech signals were manually segmented in sections (chunks) of about 3 seconds with boundaries on natural pauses within the sentences.

The total set of speech data consists of 18 matches. Chunks with speech of more than one speaker, chunks with no speech at all and very short or long chunks were discarded. The total corpus comprises 21,965 chunks, corresponding to a total duration of 10 hours. The number of chunks per match varies from 1,784 to 3,174. The number of words in the total corpus is 96,645, made up from 9,400 types.

The mean signal to noise rate (SNR) value per match varies from 6.38 to 19.09 dB. We computed the SNR of the speech per chunk. The signal energy was calculated by taking the 70% frames with the highest root mean squared energy values, and the noise energy was calculated over the remaining 30% frames. The SNR in dB is 10 times the log of the signal energy divided by the noise energy.

2.2. Recogniser

The continuous speech recogniser that was used in the experiments is Phicos [9], which is a standard HMM based system. Every 10 ms a Hamming window of 16 ms was used to compute 13 MFCCs and the log energy feature from each

frame. The first order deratives of the MFCC and logE were subsequently calculated and included in the acoustic feature vectors. In the HN experiments, mean variance normalisation was applied to the log energy features according to the method described in [11]. HN was first applied to the static MFCC and logE features, before the corresponding delta features were calculated. A set of 33 context free phone models and one non-speech model were trained. Each model consists of six states, three pairs of two identical states, one of which can be skipped. Each state consists of a maximum of 32 Gaussians. The non-speech model consists of one state.

Much effort is required to construct a good lexicon and language model (LM) for a continuous speech recognition task like this. Since this was beyond the scope of our experiments, we decided to do so called "oracle experiments" (the lexicon and LM are based on the test sets) so that recognition results are not influenced by a sub-optimal lexicon and LM. The lexicon was constructed by taking all words from the orthographic transcription of the test set. The phonetic transcriptions were made by the Institute of Phonetics of the Saarland University in Saarbrücken, Germany. The LM that was used was a combined unigram and bigram language model.

2.3. Jackknife procedure

In order to do recognition tests on the complete set of 18 matches, with independent acoustic models, a jackknife procedure was used for training and testing. To this end the total set of 18 matches was split up in three parts of six matches in such a way that the subsets were optimally balanced in terms of the number of utterances, mean SNR, playing teams, and source (video or broadcast company). Three sets of acoustic models were trained on the three possible permutations of two sets of matches. Recognition tests were done on a single match each time, using the acoustic models that were trained on the two subsets of matches, that do not include the match that is tested. E.g. the six matches from set 1 were each tested on the models trained on sets 2 and 3.

3. Histogram Normalisation

The aim of HN is to transform the test data such that the match between its overall distribution and that of the training data is improved. When HN is applied to the acoustic features used in speech recognition, it is reasonable to assume that the process which causes the mismatch has an independent effect on the different acoustic vector components. Under this assumption, each feature space dimension may be normalised independently.

The first step in performing HN is to compute the distribution of the training $(p_k(x))$ and test $(p_k(y))$ data for each feature dimension k. A cumulative distribution density is subsequently derived from both $p_k(x)$ and $p_k(y)$.

Finally, a warping function, W_k, must be derived such that:

$$P_k(x) = W_k[P_k(y)] \qquad (1)$$

HN was implemented according to the methods proposed in [4][5]. We used 128-bin histograms to approximate $p_k(x)$ and $p_k(y)$. $p_k(x)$ was calculated using all the training data while $p_k(y)$ was derived per utterance. In addition, a 3rd order spline function was used to approximate W_k. In preliminary experiments, we also investigated the possibility to estimate W_k using piece-wise linear functions. However, for short utterances the spline function estimates of W_k yielded better results than the piece-wise linear functions. The minimum and maximum values of x_k observed in $p_k(x)$ were used to limit the range of the estimation. Values in the test data that were below the minimum or above the maximum were mapped to $min(x_k)$ and $max(x_k)$, respectively.

After $p_k(x)$ was calculated from the training data, the corresponding acoustic features were also warped according to the function in Eq. (1) at utterance level. This step was taken in order to enforce training-test symmetry in terms of feature transformation. Results from similar studies have shown that the highest recognition rates are obtained if the same feature transformations are applied to the training and test data [7].

3.1. Results

For each match recognition experiments were done with and without using HN following the jackknife procedure explained in section 2.3. The results are given in WER, which was computed as follows.

$$WER\ (\%) = 100 * (\#ins + \#del + \#sub) / \#total \qquad (2)$$

where #ins is the number of inserted words in the recognition, #del the number of deleted words, #sub the number of substituted words and #total the total number of words in the test set.

	Baseline	HN
WER (%) ± 95% conf. interval	45.9 ± 0.3	43.5 ± 0.3

Table 1. Recognition scores of complete test set in WER, with and without HN

Table 1 shows the results in terms of WERs computed over the complete test set of all matches. It can be seen that applying HN significantly improves recognition performance.

In figure 1 the WERs of the baseline and HN experiments of 13 matches are compared with respect to the mean SNR of the corresponding sound files. Five matches are left out of the picture for clarity reasons (their SNR was very close to that of another match), but they contain similar results. The figure shows that for the matches with a mean SNR higher than 12 dB, HN deteriorates the recognition performance and for matches with a mean SNR lower than 12 dB, HN improves recognition performance. SNR can be seen as a measure of the extent to which the energy distributions of the noise and the speech in the data are separated. If the quality of the data is relatively high (SNR higher than 12 dB), the energy distribution of the data is bi-modal. At lower SNRs, the distribution of the energy in the data will tend to be more uni-modal. If the training data has a SNR that is much lower than the test data -the mean SNRs of the three train sets were between 12 and 13 dB-, the application of HN will cause the

original bi-modal energy distributions in the data to become more uni-modal. As a result of this data transformation, it will be more difficult for the recogniser to discriminate between speech and noise and one would expect recognition performance to deteriorate. On the other hand, if the SNR of the training data is higher than the SNR of the test data, the application of HN could improve the separation between the noise and speech components in the energy distribution of the data, thus improving recognition performance. Thus the decision whether or not to apply HN in the transcription of the MUMIS speech should be made dependent of the SNR.

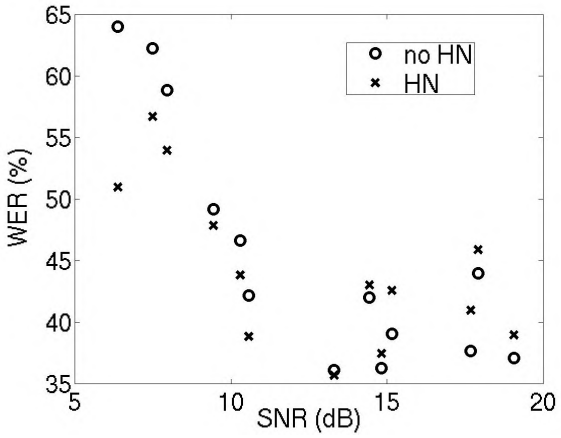

Figure 1. *WER of 13 matches and their corresponding SNR, with and without HN*

4. Ontology words and player names

From the ontology of the football domain constructed in MUMIS, a list of all terms was taken. The terms are in the categories dates, players, teams, officials, public, artefacts, body-parts, areas, spots and scenes of actions, scores, levels of competitions, events and relations. For example, 'season' is a date, 'midfielder' is a player and 'goalpost' is an artefact. In total there are 352 terms of which 131 are used in the spoken commentaries. The ontology words that are not used are typically words that do not refer to events in a match that is still ongoing. The ontology contains 59 terms that consist of more than one word, but only eight of these occur in the corpus. These multi-word expressions are treated as one word in the corpus, lexicon and language model. The mean length of the words in the ontology is 5.82 phonemes, whereas the mean length of the words in the complete test set is 4.73 phonemes. The ontology words cover only 3.0% of the total corpus.

The corpus contains 280 names of players, referees and coaches. Players are mentioned either by their last names or by the combination of their first and last names. Names existing of multiple parts are treated as one word in the corpus, lexicon and LM. The mean length of names is 7.13 phonemes and the names cover 8.2 % of the corpus.

4.1. Results

The results of the recognition experiments are presented in %correct, which indicates how well the speech recognition performs, and Category-WER (CWER), which is a measure of the quality of the input to the information retrieval system of the specific category (names, ontology words or all words).

$$\%correct = 100 * (\#sub + \#del) / \#total \quad (3)$$

$$CWER\ (\%) = 100 * (\#ins + \#del + \#c\text{-}sub) / \#total \quad (4)$$

where #ins is the number of inserted category words in the recognition, #del the number of deleted category words and #total the total number of category words in the test set. #c-sub is the number of category words that are recognised erroneously plus the number of other words that are recognised as category word. Since the latter number is independent of the denominator in the equation (the number of category words in the test set), the CWERs of different categories are difficult to compare to each other.

	%corr baseline	%corr HN	CWER% baseline	CWER% HN
All Words	55.0	58.1	45.9	43.5
Names	77.7	77.9	48.7	41.7
Ontology words	69.8	73.6	53.3	49.2

Table 2. Recognition scores of all words, names and ontology words in complete test set in terms of %correct and CWER, with and without HN

Table 2 shows that the %correct scores of the names and ontology words are higher than the corresponding scores for all words. One reason for this is that the names and ontology words are longer (on average) than all words and long words have a lower confusability. Another reason is that because the ontology words and names pronounced more clearly by the commentator then the average words, because they are important for the domain. This seems to have a positive effect on recognition results. However, the insertions and other words recognised as ontology words or names cause the CWERs for ontology words and names to be very high.

The number of insertions in the overall recognition is very low compared to the number of deletions, although the word entrance penalty is set to a very low value. This can be explained by the noisiness of the data; many words that are in recordings with a high level of background noise are recognised as noise only.

Table 2 also shows that HN seems to have more effect on names and ontology words than on the rest of the words with respect to the CWER.

5. Error Analysis

A more detailed analysis of the recognition errors in the MUMIS corpus led to the following observations that will quite probably generalize to other similar recognition tasks.

5.1. Source of the speech data

It appeared that the WERs of matches that were recorded on video (58.6% on average) were all well above the overall WER (45.9%). We computed long term average spectra (LTAS) of all matches to inspect the characteristics. The five matches from video have in common a high energy level in the frequency bands between 2.5 and 4 kHz. Figure 2 shows a typical LTAS from matches recorded in the studio and

recorded in the home. Apparently, the SNR in these bands is very low. If the speech energy in these frequency bands is well below the noise level, this constitutes a loss of information that cannot be repaired by Histogram Normalisation or cepstral subtraction.

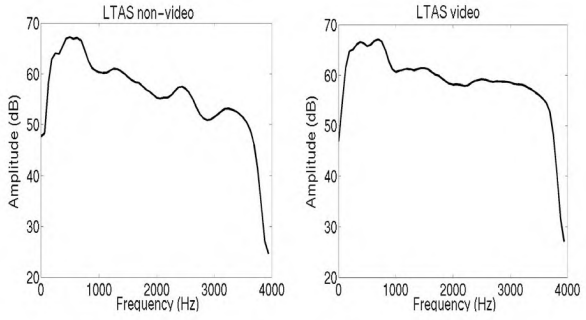

Figure 2. *LTAS of a match from the TV company (left) and one from video tape (right)*

5.2. Foreign names

From other studies [e.g. 2], it is known that foreign names tend to be recognised poorly. In the MUMIS data however, the recognition scores in terms of percentage correct for German names are comparable to those of foreign names. Taking a closer look at the recognition of names, we found that the names of German and Dutch players cause more insertion errors (2.9% and 2.4%, resp.) than names of players from other countries (avg. 0.9 %). This is explained by the fact that German and Dutch names are phonetically closer to German words than the names of players from other countries, causing a higher confusability between German and Dutch names and German words.

6. Conclusions

In this study we investigated the automatic transcription of spoken commentaries of football matches. Because the speech data contains a high level of noise, we used a technique that is known to enhance automatic recognition of noisy speech, i.e. HN The application of HN significantly improved (overall) recognition performance. It turned out that for data with a SNR lower than that of the training data, HN improved recognition performance, whereas for data with a SNR higher than that of the training data, the results were below the baseline. This suggests that it is probably counterproductive to apply HN to speech with a SNR that is better than the training data.

We also looked at the performance of the ontology words and names, because these are of crucial importance for the information retrieval task. It appeared that ontology words and names are recognised relatively better than the rest of the words. However, for the information retrieval task, the performance is disappointing.

In our error analysis, we found that loss of information in the high energy regions may explain the poor recognition results for matches recorded on video compared to the recordings of matches provided by the German TV company. Finally we saw that German and Dutch names are inserted relatively often in the German commentaries.

7. Acknowledgements

We would like to thank Loe Boves and Janienke Sturm for their useful contributions to this paper.

8. References

[1] L. Boves and N. Oostdijk, "Spontaneous Speech in the Spoken Dutch Corpus", *Proceedings ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR) 2003, Tokyo, Japan.*

[2] N. Cremelie and L. ten Bosch, "Improving the Recognition of Foreign Names and Non-Native Speech by Combining Multiple Grapheme-to-Phoneme Converters", *Proceedings ISCA ITRW Workshop Adaptation Methods For Speech Recognition 2001, pp. 151-154, Sophia-Antipolis, France..*

[3] T. Declerck, P. Wittenburg and H. Cunningham, "The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment", *Proceedings of the ACL/EACL Workshop on Human Language Technology and Knowledge Management 2001, pp 129-136, Toulouse, France.*

[4] S. Dharanipragada and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition", *Proceedings of ICSLP 2000, pp. 556--559, Beijing, China.*

[5] F. Hilger and H.Ney, "Quantile based histogram equalisation", *Proceedings of Eurospeech 2001, pp. 1135--1138, Aalborg, Denmark.*

[6] F. de Jong and Th. Westerveld, "MUMIS: Multimedia Indexing and Searching", *Proceedings of the Content-Based Multimedia Indexing workshop (CBMI) 2001, pp 423-425, Brescia, Italy.*

[7] S. Molau, M. Pitz and H. Ney, "Histogram normalisation in the acoustic feature space", *Proceedings of ASRU 2001, Trento, Italy.*

[8] D. Reidsma, J. Kuper, T. Declerck, H. Saggion and H. Cunningham, "Cross document annotation for multimedia retrieval", *Proceedings of the 3rd Workshop on NLP and XML 2003, pp 41-48, Budapest, Hungary.*

[9] V. Steinbiss, H. Ney, R. Haeb-Umbach, B. Tran, U. Essen, R. Kneser, M. Oerder, H. Meier, X. Aubert, C. Dugast and D. Geller, "The Philips research system for large-vocabulary continuous-speech recognition", *Proceedings of Eurospeech 1993, pp. 2125-2128, Berlin, Germany*

[10] J. Sturm, J.M. Kessens, M. Wester, F. de Wet, E. Sanders & H. Strik, "Automatic Transcription of Football Commentaries in the MUMIS Project", *Proceedings of Eurospeech 2003, Geneva, Switzerland.*

[11] O. Viikki and K. Laurila "Cepstral domain segmental feature vector normalization for noise robust speech recognition", *Speech Communication, 25:133-147, 1998*

[12] M. Wester, J. M. Kessens en H. Strik, "Goal-directed ASR in a multimedia indexing and searching environment (MUMIS)", *Proceedings ICSLP 2002, Denver, USA.*

[13] F. de Wet, J. de Veth, B. Cranen and L. Boves, "Additive background noise as a source of non-linear mismatch in the cepstral and log-energy domain", *Submitted to Computer, Speech and Language. 2003*