Recognizing Broadcast Conversational (BC) speech data is a difficult task, which can be regarded as one of the major challenges beyond the recognition of Broadcast News (BN).

This paper presents the automatic speech recognition systems developed by RWTH for the English, French, and German language which attained the best word error rates for English and German, and competitive results for the French task in the 2010 Quaero evaluation for BC and BN data. At the same time, the RWTH German system used the least amount of training data among all participants.

Large reductions in word error rate were obtained by the incorporation of the new Bottleneck Multilayer Perceptron (MLP) features for all three languages. Additional improvements were obtained for the German system by applying a new language modeling technique, decomposing words into sublexical components.

Index Terms: automatic speech recognition, multilayer perceptrons

1. INTRODUCTION

Concerning automatic speech recognition (ASR), the Quaero research project1 aims at developing speech recognition systems for European languages, capable of transcribing audio data available on the web. The challenge here lies in the variability of the audio data: While much experience has been gained for the recognition of broadcast news (BN) data over the last years, little progress has been reported on recognizing broadcast conversational (BC) speech as found in podcasts containing vivid discussions and frequent speaker changes.

This problem is addressed by the yearly evaluations organized within the Quaero project, every time increasing the fraction of BC speech included in the test data (50% BN and 50% BC as of 2010). In this paper the contributions by RWTH for English, French, and German are presented, leading to competitive results for all three languages and the best systems for English and German in terms of word error rate (WER), while still using the least amount of acoustic and LM training data for the German system among all partners.

All RWTH systems rely on the same basic acoustic and language modeling techniques, whereas language-specific approaches (e.g. for pronunciation modeling) are used as well as sharing of training results across languages (as for MLP features) whenever appropriate to optimize systems and reduce computational effort.

The rest of this paper is organized as follows: Section 2 gives details on the acoustic modeling including MLP features and the data that were used for acoustic training. Section 3 describes the language modeling for standard and sublexical-based methods, followed by section 4 explaining details of the pronunciation modeling for the RWTH systems. Segmentation and recognition are covered in subsequent sections. In section 7, experimental results are discussed. Section 8 concludes the discussion.

2. ACOUSTIC MODELING

2.1. Baseline acoustic modeling

For all languages, several subsystems were trained, based on the Mel-Frequency Cepstral Coefficients (MFCC) using a bank of 20 filters. For each time frame 16 coefficients including energy were extracted and a cepstral mean and variance normalization was applied on the segment level.

Augmenting the MFCC features by a voicedness feature and applying a sliding window of size 9, 154-dimensional feature vectors were obtained that were projected down to 45 components using an LDA transformation. To introduce variability between the resulting subsystems and thereby improve the final system combination step, Perceptual Linear Predictive (PLP) features were also extracted in an analogous manner. Then phone-posterior-based features, estimated using a multilayer perceptron (MLP), were appended. The next subsection gives more details about those features part of the tandem approach where conventional short-term features and MLP features are combined in a single system.

---

1http://www.quaero.org
Table 2. MLP feature training data and network topology

<table>
<thead>
<tr>
<th>Training data</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR 1</td>
<td>Quaero</td>
</tr>
<tr>
<td>FR 2</td>
<td>Quaero, Ester1+2</td>
</tr>
<tr>
<td>EN 1</td>
<td>HUB4+TDT4</td>
</tr>
<tr>
<td>EN 2</td>
<td>HUB4+TDT4</td>
</tr>
<tr>
<td>DE 1</td>
<td>344 h BC+BN+read speech</td>
</tr>
<tr>
<td>DE 2</td>
<td>Quaero</td>
</tr>
</tbody>
</table>

Table 3. Language model training data for each language

<table>
<thead>
<tr>
<th>Corpus</th>
<th># running words</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN Gigaword</td>
<td>2.6B</td>
<td>newspaper</td>
</tr>
<tr>
<td>Quaero</td>
<td>461 M</td>
<td>blog+news</td>
</tr>
<tr>
<td>Quaero</td>
<td>4.6 M</td>
<td>BC+BN</td>
</tr>
<tr>
<td>FR Gigaword</td>
<td>837 M</td>
<td>newspaper</td>
</tr>
<tr>
<td>Quaero</td>
<td>262 M</td>
<td>blog+news</td>
</tr>
<tr>
<td>Web data</td>
<td>248 M</td>
<td>archives+news</td>
</tr>
<tr>
<td>transcriptions</td>
<td>3 M</td>
<td>BC+BN</td>
</tr>
<tr>
<td>DE Quaero</td>
<td>228 M</td>
<td>blog+news</td>
</tr>
<tr>
<td>Web data</td>
<td>306 M</td>
<td>archives+news</td>
</tr>
<tr>
<td>transcriptions</td>
<td>0.5 M</td>
<td>BC+BN</td>
</tr>
</tbody>
</table>

2.2. MLP features

Showing significant improvements, MLP features for acoustic modeling have attracted much attention during the last years. Considerable effort has been spent on improving the input features of the neural networks as well as on optimizing the network topology itself.

Regarding the topology, competing approaches included a hierarchical processing (H-MLP) as well as the introduction of a hidden ‘bottleneck’ layer. This layer corresponds to a dimensionality reduction, as not the outputs of the network are used as tandem features directly but the outputs of this lower-dimensional hidden layer. Both ideas were successfully combined in [1], resulting in the hierarchical bottleneck features (HBN-MLP). For our systems, both H-MLP and HBN-MLP features were used.

The neural networks were trained by feeding the Multiresolutional RASTA (MRASTA) features as inputs and the phone-posterior probabilities, computed by a forced alignment of the acoustic training data, as desired outputs. As a last step, the MLP features were decorrelated by a PCA transformation. This also allows an additional dimensionality reduction in case of the H-MLPs. Table 2 lists the different MLP feature types and the corresponding acoustic data that were used for training.

Subsystems also improved when exchanging MLP features across languages (though estimated on different phone sets). This property, first reported in [2], was made use of to increase subsystem variability.

2.3. Speaker normalization and adaptation

For most of our subsystems, MFCC and PLP features were normalized using Vocal Tract Length Normalization (VTLN). Recent experiments showed that although the gain induced by VTLN is mostly included in that obtained by CMLLR adaptation, there may still remain minor improvements of up to 0.2% absolute in WER, and no deterioration was ever observed.

2.4. Discriminative training

To sharpen acoustic models, discriminative training was applied. Lattices were computed using the current best acoustic models for each language, where only a unigram language model was used to avoid overfitting. Based on these lattices, the Minimum Phone Error (MPE) training criterion was optimized, [4].

3. LANGUAGE MODELING

Based on the available training data, 4-gram language models (LMs) were estimated for each language using [10], smoothed by the Modified Kneser-Ney method. We partitioned the LM data into blocks, estimating n-gram probabilities for each block individually. Then the LMs were linearly interpolated while optimizing the perplexity on a holdout data set.

Table 3 gives an overview of the training material that was used. For the 2010 evaluation, substantial amounts of in-domain text data downloaded from web blogs were distributed to all participants. Because of the comparatively small-sized Gigaword corpus for French and the lack of a Gigaword corpus for German, the LM data were extended by text resources from the web by crawling RSS news feeds and web archives for these two languages.

A high lexical variety is characteristic for the German language, as a large number of distinct lexical forms can be generated by derivation, compounding, and inflection. For this reason, an alternative LM approach based on word decomposition into sublexical fragments was investigated to reduce OOV rates and minimize perplexity. This method was shown
to have good performance in Arabic which also is a morphologically rich language.

Words were decomposed using a statistical data-driven tool. Then a standard \( n \)-gram LM was estimated on the decomposed text. To prevent confusion between the most important German words and other less-frequent fragments, the most frequent decomposable words were excluded from decomposition as proposed in [5]. Finally perplexity values of the LMs were 269, 130, and 131 for German, English, and French, respectively. In spite of the huge vocabulary size for German (300K), the OOV rate was still at 1.13 %, while for English and French it was around 0.5 % for vocabulary sizes of 150K and 200K only.

4. PRONUNCIATION MODELING

The recognition of BC data also necessitates updated and improved pronunciation lexica, as speakers tend to pronounce words less carefully in conversational speech.

In case of the German system, this problem was addressed by introducing automatically generated pronunciation variants into the training lexicon. A large number of pronunciation variants was derived from those found in the baseline lexicon in a systematic manner by the deletion of phonemes. After computing a forced alignment using the updated lexicon, only those pronunciations were retained for future training and recognition that were seen in training at least a certain number of times. By this technique the WER was reduced from 21.4 % by 0.4 % absolute.

For the English system, pronunciations were improved manually for the most frequent words, also adding additional pronunciation variants and training pronunciation weights which gave minor improvements in WER.

For French, a new lexicon was set up based on the widely used BDLex. The advantage of this baseline lexicon compared to other resources lies in its thorough coverage of different pronunciation variants including those induced by French-specific phenomena like liaison and schwa-deletion. As this results in a huge blow-up of lexicon entries, additional meta-phonemes representing a set of several normal phonemes were introduced to reduce the number of pronunciation variants. In combination with pronunciation weights, the WER was improved from 24.4 % by 0.4 % absolute.

For all languages, grapheme-to-phoneme models were trained for the creation of pronunciations not found in the baseline lexica, as described in [6].

5. SEGMENTATION

The audio segmentation of the RWTH system makes use of a log-linear classifier that decides if a time frame corresponds to a segment boundary or not. The features for the log-linear model were chosen as to cover e.g. the variability of the acoustic signal, the speaker homogeneity, and also changes in the acoustic conditions.

In addition, the set of log-linear features was augmented by information obtained by a one-pass recognition on the unsegmented audio data. This includes the number of words within a segment as well as the time stamps of sentence boundary tokens hypothesized by the recognizer.

6. RECOGNITION SETUP

The RWTH systems rely on five subsequent recognition passes, as depicted in Figure 1. In an initial unadapted pass, a first transcription was obtained which formed the basis for the second, CMLLR-adapted recognition pass.

The resulting transcriptions then were exchanged between subsystems for cross-adaptation, leading to the third full recognition pass. In case of English and French, for cross-adaptation only the previous recognition output of one subsystem was used, whereas for German, all subsystems were used for CMLLR-adaptation as indicated by the dashed lines.

As for the full recognition only a pruned LM was used, third pass lattices were rescored using the full LM. System combination was applied as a last step. The lattices were converted to a confusion network (CN) by an iterative procedure. Afterwards, the final transcription was obtained by CN combination as presented in [8]. In case of the German system, a post-processing step was also added to concatenate numbers to avoid spelling errors.

7. EXPERIMENTS

Table 4 shows detailed WER results for the individual subsystems of all languages. For English, also the different acoustic training corpora for the individual subsystems are given. For the other languages, these data were kept fixed, always using all material from Table 1.

The results from the table show that the new HBN-MLP features gave a significant gain in WER compared to the H-MLP features, as can be seen e.g. in the performance of the German \( s_1 \) and \( s_3 \) systems. Incorporating HBN-MLP features improved an English speaker-adapted ML-trained MFCC system by 18 % relative, including a gain of 3 % relative compared to the older H-MLP features.
Furthermore, it seems that the quality of the MLP features strongly depends on the availability of large amounts of training data. Regarding the French system, the English H-MLP features eventually led to better performance than the more-advanced HBN-MLPs that were trained on only 80 h of in-domain French training data.

Additional conclusions can be drawn from the relative improvements as summarized in Table 5 that also gives baseline results for the 2009 systems ([11]). For all languages, gains by MPE were quite small which we assume to be caused by the incorporation of MLP features that already form a discriminative training approach.

For the German language, the sublexical language modeling approach yielded an improvement from 22.5 % to 21.7 % in WER after CMLLR adaptation.

All systems could be improved significantly compared to last year’s systems which was also due to the availability of more in-domain training data distributed to all evaluation participants.

Table 4. Results on the 2010 development set, STF = short-term features, last line showing results for the 2010 evaluation data (numbers in parentheses give Character Error Rates)

<table>
<thead>
<tr>
<th></th>
<th>DE (s1)</th>
<th>EN (s1)</th>
<th>FR (s1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>26.2 %</td>
<td>33.0 %</td>
<td>32.2 %</td>
</tr>
<tr>
<td>VTLN</td>
<td>21.9 %</td>
<td>22.4 %</td>
<td>25.5 %</td>
</tr>
<tr>
<td>+CMLLR</td>
<td>20.2 %</td>
<td>19.8 %</td>
<td>23.2 %</td>
</tr>
<tr>
<td>+MPE</td>
<td>19.9 %</td>
<td>19.4 %</td>
<td>22.6 %</td>
</tr>
<tr>
<td>+X-adaptation</td>
<td>19.9 %</td>
<td>19.4 %</td>
<td>22.5 %</td>
</tr>
<tr>
<td>+LM-rescoring</td>
<td>19.8 %</td>
<td>19.1 %</td>
<td>22.3 %</td>
</tr>
</tbody>
</table>

Table 5. Improvements in terms of WER obtained by different methods, measured on the 2010 development corpus.

8. CONCLUSIONS

In this paper the RWTH ASR systems for English, French, and German were presented in detail which took part in the 2010 Quaero evaluation for BC and BN data. Compared to last year, huge improvements in WER were obtained. Contributing to these results were advances in MLP feature design, notably the new Hierarchical Bottleneck MLP features. For German, a new language modeling approach was tried that showed a significant gain in WER which is consistent with experiments for different morphologically rich languages like Arabic. Further improvements were related to pronunciation modeling, systematically creating new pronunciation variants, and the training of the segmentation software, adapting the models to the BC+BN task. Thanks to the improvements, RWTH obtained the best WERs for English and German.

Acknowledgement—This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

9. REFERENCES


