SHoUT, the University of Twente Submission to the N-Best 2008 Speech Recognition Evaluation for Dutch

Marijn Huijbregts, Roeland Ordelman, Laurens van der Werff and Franciska de Jong

Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands

{huijbreg,ordelman,werff,fdejong}@ewi.utwente.nl

Abstract

In this paper we present our primary submission to the first Dutch and Flemish large vocabulary continuous speech recognition benchmark, N-Best. We describe our system workflow, the models we created for the four evaluation tasks and how we approached the problem of compounding that is typical for a language such as Dutch. We present the evaluation results and our post-evaluation analysis.

Index Terms: LVCSR, Benchmark

1. Introduction

In 2006 a research project called N-best\(^1\) was started [1]. N-Best aims at setting up the infrastructure for a benchmark evaluation in large vocabulary speech recognition for the Dutch language, and at conducting the evaluation. This evaluation, the first one for Dutch ASR, focuses on two tasks: broadcast news and Conversational Telephone Speech (CTS). Within these tasks, speech recognition performance for both Northern and Southern Dutch as it is spoken in the Netherlands and Flanders (Belgium) respectively, are evaluated. In this paper the N-Best submission of University of Twente is presented and discussed.

In previous years, we have developed a Large Vocabulary Continuous Speech Recognition (LVCSR) system called SHoUT\(^2\). Although parts of this system have been tested in other benchmarks, the N-Best project provided us the opportunity to benchmark the entire Dutch system for the first time and compare it to the competing systems.

In this paper, after discussing our system workflow (section 2), we will focus on how we created specific models for the two dialects Dutch and Flemish (section 3) and how we approached the compounding problem typical for languages such as Dutch (section 4). Section 5 presents the benchmark results followed by a post-evaluation analysis in section 6.

2. System description

In this section we will briefly describe our four N-Best submissions, but first we will discuss the general workflow of SHoUT, our LVCSR system.

2.1. The SHoUT LVCSR system workflow

The workflow of the SHoUT LVCSR system is depicted in figure 1. Processing of an audio file starts with speech activity detection (SAD) in order to filter out the audio parts that do not contain speech. The SAD module was evaluated earlier at the NIST Rich Transcription benchmark in 2007 (RT07s[2]) and described in-depth in [3].

After SAD, the speech fragments are segmented and clustered. In this step, the speech fragments are split into segments that only contain speech from one single speaker with constant audio conditions. Note that for instance speech from a telephone recording and high quality recordings are separated as well, even if the speech comes from only one speaker. Each segment is labeled with a speaker ID. For this segmentation and clustering step, the speaker diarization module presented and evaluated at RT06s [4] was used.

As shown in figure 1, the ASR module incorporates four consecutive processes. First, features are extracted from the segmented audio and normalized for speaker and audio variations. Next, a primary decoding pass is run. The output of this pass is used for adapting the acoustic model for each speaker cluster. Finally, the secondary decoding pass uses the adapted models for producing the final transcription.

For feature extraction, two techniques were used for feature normalization: Cepstrum Mean Normalization (CMN) and Vocal Tract Length Normalization (VTLN).

The ASR decoder applies a time synchronous Viterbi search. The Viterbi search is implemented using the token passing paradigm [5]. HMMs with three states and GMMs for their probability density functions are used to calculate acoustic likelihoods of context dependent phones. Up to 4-gram back-off language models (LMs) are used to calculate the priors. In [6] the decoder is described in more detail.

The clustering information obtained during segmentation and clustering is used to create speaker dependent acoustic models. The SMAPLR adaptation method from [7] is used to adapt the means of the acoustic model Gaussians. This method has the benefit that it requires hardly any tuning and automatically determines to what extent the models can be adapted given the amount of adaptation data that is available. This procedure prevents the models from being over-fitted on the adaptation data when only small amounts of adaptation data are available.
2.2. Workflow in the N-Best submissions

The N-Best evaluation consists of four tasks: Broadcast News (BN) for Dutch and Flemish, and Conversational Telephone Speech (CTS) for Dutch and Flemish. For the two BN tasks, the audio is processed according to the procedure described above: first segmentation and speaker diarization, followed by the determination of the VTLN warping factor of each speaker cluster. The first decoding iteration is performed using BN acoustic models, vocabulary and language model. After unsupervised adaptation of the acoustic models of each speaker cluster, the final decoding iteration is performed.

For the two conversational telephone speech tasks, diarization is not needed because each speaker is recorded on his own channel. Some recordings in the development set contain cross-talk, the phenomenon that a speaker is recorded on the channel of the other speaker. Speaker diarization could have been employed to remove the cross-talk, but the N-Best project guaranteed that there would be no cross-talk in the evaluation data and therefore the diarization step is omitted. Instead, for each audio channel a simple energy-based segmentation is conducted in order to obtain the speech segments. These segments are used directly for determining the VTLN warping factor of each speaker. After a first decoding pass the acoustic models are adapted and with the second decoding run the final hypothesis is generated.

The BN and CTS systems both make use of a post-processing component at the very end of the process. These steps improve the system concerning capitalization, numbering and word compounding and are described in section 5. First, in the following section we will describe our approach for creating the most optimal models for Dutch and Flemish.

3. Dialect specific models

The data that can be used to train the statistical models is defined by the N-Best organizers. In the following subsections a number of experiments will be described in which the Dutch and Flemish data is mixed in order to create optimal language models and acoustic models for both dialects.

3.1. The language models

For language modeling, in addition to speech transcripts from the Spoken Dutch Corpus (CGN) [8], Dutch and Flemish newspaper data was available: some 450M words from the Twente News Corpus (TWNCSpeech) and more than 1400M words from the Mediargus corpus respectively.

For each dialect (Flemish and Dutch), a single language model is created that is used for both the CTS and BN task. Table 1 contains the results of experiments on our Dutch BN development set with our baseline acoustic model. The results show that the mix of newspapers and speech transcriptions performs best on this set. This language model is obtained by mixing the two models that are created using newspaper text and the speech transcriptions. The same procedure is followed for Flemish, but because of the lack of development data, no experiments are conducted to determine whether mixing newspaper and transcription data is best for Flemish as well.

3.2. The acoustic models

For each evaluation condition, data for BN and CTS, both for Dutch and Flemish, are available for training the acoustic models. We have trained acoustic models for each condition using the individual data sets, but as Dutch and Flemish are dialects, performance of the system could possibly be improved by using all available acoustic data, so both Dutch and Flemish, for model training. Assuming: ’there’s no data like more data’, we trained such a model and tested it on the Dutch BN development set.

The result of this experiment (see table 2) shows that using all data (model 2) does not improve performance opposed to using dialect specific data (model 1). Apparently the pronunciation of the same phonemes by Dutch and Flemish speakers are sufficiently different to warrant the training of separate acoustic models.

<table>
<thead>
<tr>
<th>Acoustic model</th>
<th>% WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. AM based on Dutch data</td>
<td>28.1</td>
</tr>
<tr>
<td>2. AM based on all data</td>
<td>28.4</td>
</tr>
<tr>
<td>3. Mix of the first two models</td>
<td>27.8</td>
</tr>
</tbody>
</table>

Table 2: Results of experiments on the Dutch BN development set with a model trained on Dutch data, a model trained on Dutch and Flemish data and a mix of these two models.

During the training of the AM models, it was noticed that the likelihood on the training set increased for some phones and decreased for other phones when the Flemish data was added to the training set. Perhaps for some phones it actually does help to add the Flemish data and for other phones it is better not to do this. In order to test this hypothesis, the overall likelihood on the Dutch BN training data set is determined for each phone of both the AM trained on Dutch BN and the AM trained on all BN data. A third AM is then created by selecting the model with the highest likelihood for each phone. The WER on the Dutch BN development set using this mixed AM is shown in table 2 as model 3. This AM outperforms the AM trained on Dutch BN and therefore the mix AM is used in the SHoUT submission for N-Best. For the three other tasks, the same procedure is followed for training the AM.\footnote{Unfortunately, just before the submission deadline a bug was found in this procedure for CTS-VL and therefore for this task the acoustic model trained solely on Flemish CTS data is used.}

<table>
<thead>
<tr>
<th>Language model</th>
<th>PP % WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newspaper LM</td>
<td>147</td>
</tr>
<tr>
<td>Newspaper/transcriptions LM</td>
<td>131</td>
</tr>
</tbody>
</table>

Table 1: Perplexity and WER of newspaper LM and mixture LM that also incorporates speech transcripts.
4. Post-processing

The scoring method of N-Best is very strict in four aspects. First, scoring is performed case-sensitive. Also, strict rules are defined for the way numbers should be compounded. Third, a lot of compounded words exist in the Dutch language. Words such as: ‘waterpolo bal’, are written as multiple words in English (‘water polo ball’). For N-Best, all words not compounded correctly will be considered incorrect. Finally, for the N-Best evaluation it is allowed to label filled pauses such as ‘eh’ or ‘uhm’, as non-lexical so that these words will be discarded during scoring. In order to face these four rules, a post-processing component is added to the ASR module that handles the case of each word, the re-writing of numbers, compound restoration and the labeling of filled pauses.

The three words from the dictionary: ‘eh’, ‘uhm’ and ‘mmhh’, are marked as filled pause. Whenever these words are recognized, they will be labeled as such. For compound restoration, a list of possible compounds, abstracted from the Dutch newspaper corpus, is used. Whenever two words are recognized of which the compound is present in the list, the words are merged. The vocabularies are all case insensitive. For case-restoration, the disambiguate tool from the scri-lm toolkit is used [10]. Case sensitive language models are created for Flemish and Dutch, using the newspaper material. These models are used to map the lower-case recognitions to the correct case with the disambiguate tool. The Dutch and Flemish dictionaries contain all possible numbers, but not in concatenated form. It contains the numbers ‘honderd’ and ‘drie’ but not ‘honderddrie’. Therefore a script is written that concatenates all numbers according to the rules of the evaluation. In section 6, each post-processing step is evaluated.

5. Evaluation results

The evaluation results are listed in table 3. The performance of our system was competitive to the other submissions. On average, looking at all four tasks, we ranked third in the evaluation.

Two aspects of the results listed in table 3 stand out. The results of the broadcast news tasks are lower compared to the development set (For Dutch, 27.5% compared to 39.4% WER) and adaptation decreases the results for the CTS tasks. In the next section, the post-evaluation analysis, these two aspects will be investigated.

6. Post-evaluation analysis

The performance of our system on the Dutch and Flemish BN evaluation data are not as good as on the Dutch BN development set. In part, the word error rates are higher because the data contain a good amount of interviews and discussions. The development data consists mainly of prepared studio speech. In table 4 the evaluation results of the Dutch BN task are shown for the four conditions with which the evaluation data was labeled. The word error rate of the clean studio condition is 26.3%. This is comparable to the results obtained on the development set (27.5%).

Table 3: The N-Best evaluation results for each task and the significance of the difference between the first and second decoding iteration.

<table>
<thead>
<tr>
<th>Task</th>
<th>1&lt;sup&gt;st&lt;/sup&gt; iteration %WER</th>
<th>2&lt;sup&gt;nd&lt;/sup&gt; iteration %WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN-Dutch</td>
<td>41.3</td>
<td>39.4</td>
</tr>
<tr>
<td>BN-Flanders</td>
<td>34.6</td>
<td>33.6</td>
</tr>
<tr>
<td>CTS-Dutch</td>
<td>60.4</td>
<td>60.7</td>
</tr>
<tr>
<td>CTS-Flanders</td>
<td>72.1</td>
<td>72.8</td>
</tr>
</tbody>
</table>

Table 4: The N-Best evaluation results for the Dutch BN task. The results are shown for the main four audio conditions that are present in the task. The word error rate (WER) is divided into substitution (SUB), deletion (DEL) and insertion (INS) errors.

The difference in evaluation and development data is not the only problem. After studying the evaluation results, it became clear that a high amount of speech was discarded by the segmentation module that falsely labeled this speech as audible non-speech. Inspection of these segments revealed that the module filtered all speech out of the system that was recorded over a telephone line. The deletion percentage of 43.1% WER in the F2 condition in table 4 is a clear indication of this problem. The development set did not contain any telephone speech and therefore this flaw in the system was not noted before. Next, a short explanation of this problem in the segmentation module will be given.

6.1. Telephone speech in broadcast news

Our segmentation, described in-depth in [3] first identifies speech segments using a bootstrapping SAD component. With use of this initial segmentation, a speech model, silence model and audible non-speech model are generated. These models are applied in a second pass, resulting in the final segmentation.

The bootstrapping is performed by a model based speech/silence segmentation component trained solely on broadband speech. The telephone speech from the evaluation data did not fit the broadband speech model well and was classified as non-speech during the first bootstrapping run and in the second pass, all telephone speech was classified as audible non-speech.

To avoid this problem a narrow-band/broadband detection module can be applied before segmentation is performed. It is also possible to adjust the segmentation module and use speech model for both narrow-band and broadband in the first bootstrapping pass so that it is more robust for this channel problem.

Table 5: The N-Best evaluation results for the Dutch BN task, where segments originally labeled as audible non-speech, are processed by the ASR module using the telephone acoustic models. The significance levels are measured compared to the original submission (table 3).

<table>
<thead>
<tr>
<th>Task</th>
<th>BN model %WER</th>
<th>CTS model %WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN-Dutch</td>
<td>35.5</td>
<td>34.9</td>
</tr>
<tr>
<td>BN-Flanders</td>
<td>33.5</td>
<td>31.7</td>
</tr>
</tbody>
</table>

For this specific evaluation, that guarantees that the audio
fragments do not contain any audible non-speech, it is possible to interpret the results of the segmentation module slightly differently. In this case, all audio that is used for training the audible non-speech model is known to contain high energy levels and mismatch the acoustical conditions of the training data. The high energy levels indicate that the fragments are not silence and therefore must be speech. The obvious condition that does not match the training data is speech over telephone lines and therefore the segmentation results can be interpreted as a silence/studio-speech/telephone-speech classification. Table 5 contains the results of experiments on the Dutch and Flemish BN task where the audible non-speech segments were interpreted as being telephone speech. These segments are passed to both the broadcast news ASR module and the CTS ASR module⁴. The experiments show that indeed the best results are obtained by applying the CTS acoustic models. The word error rate of the telephone condition (F2) for the Dutch BN task is 39.6% (8.5% WER deletions) when using the CTS models.

6.2. Post-processing

In section 3, the post-processing steps were described that are added to the system for the N-Best evaluation. Table 6 contains the word error rates after each post-processing step for the Dutch BN task (where the telephone speech is decoded with the CTS acoustic models). Baseline performance without post-processing results in 38.6% error rate. Table 6 shows how each of the steps improves on this result. Although the contribution of some steps is marginal, all improvements are significant with \( p < 0.001 \).

<table>
<thead>
<tr>
<th>Post-processing step</th>
<th>%WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>No post-processing, scored case-sensitive</td>
<td>38.6</td>
</tr>
<tr>
<td>Filled pauses</td>
<td>37.9</td>
</tr>
<tr>
<td>Filled pauses and compounds</td>
<td>37.6</td>
</tr>
<tr>
<td>Filled pauses, compounds and case</td>
<td>35.3</td>
</tr>
<tr>
<td>Filled pauses, compounds, case and numbers</td>
<td>34.9</td>
</tr>
<tr>
<td>All post-processing, scored case-insensitive</td>
<td>33.8</td>
</tr>
</tbody>
</table>

Table 6: Results of post-processing experiments on the N-Best Dutch BN evaluation data. All experiments are significant with \( p < 0.001 \).

If the task is scored case-insensitive, the WER is reduced with almost one percent. Although the case of the reference transcription is not correct in all places, this means that the case normalization step shows room for improvement.

7. Discussion

The results of the N-Best evaluation show that on clean broadcast news speech, the system performance is comparable to the results during development. The WER on the spontaneous speech and the speech under degraded conditions is considerably higher.

The evaluation revealed a weak spot of the segmentation module. Although the behavior of the module when segmenting BN audio that contains telephone speech is logical, it was not expected. It is possible to solve the problem by adding a broadband/narrowband classification module, but it is preferred to adjust the segmentation module so that it is possible to detect speech with various conditions that are all represented by bootstrap speech models. With the current module it is only possible to detect the broadband speech represented by the bootstrap BN model, but it should be possible to add other models such as the CTS model during the bootstrapping step.

Compared to the SHoUT vocabularies, some participants of the N-Best evaluation used vocabularies with a considerably higher number of words. A vocabulary with 500K words was even used that has an out-of-vocabulary rate on the development set of only 0.5%. It would be interesting to test if the SHoUT performance can be improved by increasing the number of words in the vocabularies.

8. Acknowledgments

The work reported here was supported by the BSIK-program MultimediaN which is funded by the Dutch government under contract BSIK 03031, by the European Union 6th FWP IST Integrated Project AMIDA (Augmented Multi-party Interaction with Distant Access, FP6-506811) and the project CHoral - Access to Oral History, which is funded by the NWO program CATCH (http://www.nwo.nl/catch).

9. References


⁴For this experiment, only one decoding pass is used. The models are not adapted for a second iteration.