A Broadcast News Corpus for Evaluation and Tuning of German LVCSR Systems

Felix Weninger[1]*, Björn Schuller*[1]*, Florian Eyben*[1]*, Martin Wöllmer*[1]*, and Gerhard Rigoll*[1]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
{weninger|schuller|eyben|woellmer|rigoll}@tum.de

Abstract

Transcription of broadcast news is an interesting and challenging application for large-vocabulary continuous speech recognition (LVCSR). We present in detail the structure of a manually segmented and annotated corpus including over 160 hours of German broadcast news, and propose it as an evaluation framework of LVCSR systems. We show our own experimental results on the corpus, achieved with a state-of-the-art LVCSR decoder, measuring the effect of different feature sets and decoding parameters, and thereby demonstrate that real-time decoding of our test set is feasible on a desktop PC at 9.2 % word error rate.

Index Terms: LVCSR, corpus, German, broadcast news, real-time

1. Introduction

Transcription of broadcast news has been an active topic of research in the last decade. Various languages apart from English [1, 2] have been covered, including Arabic [3], French [4], German [5, 6], and Mandarin [7, 8]. The process of full transcription is usually divided into a segmentation and speech recognition step: first, the audio signal is segmented into speech and non-speech parts (e. g. music), and then the speech parts are decoded by a speech recogniser [9].

In this paper, we do not aim at a full-featured transcription system, but rather focus on the challenge of speaker-independent German large-vocabulary continuous speech recognition (LVCSR), using a broadcast news speech corpus for evaluation. This corpus was originally created at Duisburg University in the context of the ALERT project, and used as a whole for training a broadcast news transcription system [5]. In that study the test set was rather small and not clearly specified. In contrast, this paper proposes a suitable corpus structure for systematic tuning and evaluation of a German LVCSR system, including subdivision into training, development, and test sets. We believe that provision of the corpus will be a valuable contribution for the engineering of speech recognition tools for diverse applications in the future, since recognition of German broadcast news is a considerably challenging application for LVCSR: on the one hand, recognition of broadcast news has to utilise large dictionaries to avoid frequent out-of-vocabulary events; on the other hand, it has to deal with the pecularities of the German language, such as frequent inflection (e. g. German *einen* or *einem* = English *a(n)*) and complex compound words (e. g. German *Spracherkennungsproblem* = English *speech recognition problem*) [6].

Our work focused on structuring and preparing the German speech data available from the ALERT project into a corpus suitable for evaluation of speaker-independent LVCSR systems (Section 2), assessing the benefit of different acoustic feature sets (Section 3), and fine-tuning of the decoding parameters (Section 4) with respect to the trade-off between real-time factor (RTF) and word error rate (WER). We will present the results of our experiments in Section 5, and show that real-time processing on a state-of-the-art PC is feasible at a WER of less than 10 %, which is lower than the ones typically found in previous works on transcription of German broadcast news.

2. The Broadcast News Corpus

The Broadcast News (BCN) Corpus was recorded in 2000–01 at Duisburg University and consists of over 160 hours of German speech, including mainly radio broadcasts, but also news on television (*Tagesschau*). For the television parts, video is available as an additional modality, e. g. for audiovisual speech recognition, but was not considered in our experiments. Audio is sampled at 16 kHz with 16-bit PCM encoding. The audio files each correspond to a broadcast with several minutes length. The segmentation and transcription is contained in a flexible XML file format that is illustrated in Figure 1: first, each of the audio files is coarsely divided into *sections*, corresponding e. g. to jingles at the beginning of the broadcast, or speech parts. Each speech part is segmented into *turns* of different speakers which are listed at the beginning of the file. Gender information could be used to train specific acoustic models, which we leave as a topic for further experiments. Each turn corresponds to one or several news messages read by one speaker and is transcribed on the word level, using reference texts from the web sites of the radio and TV stations whenever available. Manually created alignments (*sync times*) divide the turns into *utterances*, which usually correspond to complete news messages and hence contain a few sentences. The average utterance length in the corpus is 80 words. Alignments are also provided whenever the background condition (clean speech, music, background noise, or a combination of them) changes, which allows further partitioning of the utterances into segments with only a specified background condition. Finally, signal segments with low speech quality – due to transmission or microphone problems – are also identified. We divided the utterances from the clean speech parts into a training, development, and test set, ignoring the low-quality segments (approx. 1.4 % of the total utterance length). We realised the constraint of speaker-independency by choosing these sets such that no radio or TV channel occurs in more than one set. The recording length and number of utterances for the three sets is shown in Table 1, while the number of out-of-vocabulary (OOV) entries is outlined in Table 2 for a 148 K language model vocabulary size.

3. Feature Extraction

To be able to measure how recognition performance is affected by different acoustic features, we extracted three feature sets.

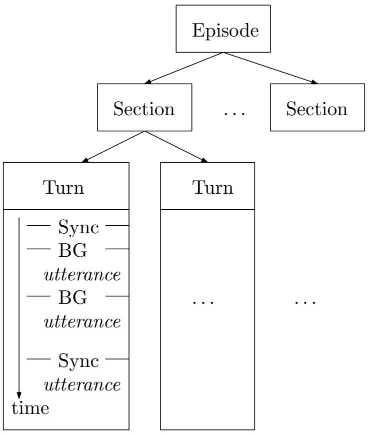

Figure 1: *Hierarchical structure of the XML format. It allows a flexible annotation and segment extraction from audio files (*Episode*) by providing* Sections *corresponding e. g. to speech or music parts, subdivided into* Turns *of a certain speaker, which consist of a sequence of utterances for which time alignments (*Sync*) are provided. Changes in background (*BG*), such as noise conditions are also annotated.*

Channel	**Type**	**Year(s)**	**h**	**# Words**
Training set				
DLF	Radio	2000–01	67.81	535 K
SWR2	Radio	2000	38.42	333 K
SWR2 Info	Radio	2000	1.88	18 K
WDR2	Radio	2000	26.75	224 K
NDR4	Radio	2000	11.41	98 K
total			146.28	1 209 K
Development set				
ARD	TV	2000–01	8.93	84 K
Test set				
DW	Radio	2000–01	6.33	50 K

Table 1: *Overview over the training, development, and test set of the BCN corpus. "h" indicates total utterance length in hours. Note that for our experiments we only used the audio channel of the TV data.*

A baseline set contained Mel Frequency Cepstral Coefficients (MFCC) 1–12 and signal log-energy, and their first (δ) and second order regression coefficients ($\delta\delta$), which were extracted using the openEAR feature extractor [10], and are identical to the features extracted by the HTK toolkit [11].

Next, this feature set was augmented by 12 Line Spectral Pairs (LSP), along with their δ and $\delta\delta$ coefficients, computed with openEAR. To reduce the number of parameters that had to be estimated for the acoustic models, and to have feature sets of comparable dimension, we applied Principal Components Analysis (PCA) to the 75-dimensional feature vectors. For PCA, we computed the covariance matrix of the training set when testing with the development set, or the one of the union of training and development set when testing with the test set. The feature space was reduced by considering only the first N eigenvalues (in descending magnitude) $\lambda_1, \dots, \lambda_N$ such that

$$\frac{\sum_{i=1}^{N} \lambda_i}{\sum_{i=1}^{75} \lambda_i} \geq 0.9999, \qquad (1)$$

Set	**# OOV**	**% OOV**
Training	3 585	0.55
Development	734	1.11
Test	127	0.36

Table 2: *Number of unique out-of-vocabulary entries* (# OOV)*, and out-of-vocabulary rate* (% OOV)*, related to the 148 K language model vocabulary size, in the training, development, and test set of the BCN corpus. Note that we handle OOV words in the training set by mapping to a garbage model.*

and choosing the corresponding eigenvectors as basis. For both cases (training set or union of training and development set), the resulting dimensionality was $N = 38$. The final features will be denoted by MFCC-LSP-PCA.

Finally, as a third feature set, we chose the Perceptual Linear Prediction (PLP) coefficients 1–12 along with log-energy and their δ and $\delta\delta$ coefficients, computed with HTK [11].

4. Acoustic and Language Models

In contrast to previous work on German broadcast data [5] using the "Ducoder" stack decoder [12], we implemented a LVCSR system on top of HDecode [11]. Our acoustic models included tied-state Hidden Markov Models (HMMs) with three emitting states, corresponding to decision-tree clustered cross-word triphones built from 54 monophone models. Apart from 51 models for the common German phoneme alphabet and silence, the monophones also included a model for the nasal *en* vowel that quite frequently occurs in German due to a considerable number of French foreign words (e. g. *Arrangement*). Two additional models were needed for technical reasons: a 'garbage' phoneme model was trained to model out-of-vocabulary (OOV) words in the training set, in order to be able to use all available utterances from the training set. Also, due to utterances containing multiple sentences in the corpus, it was necessary to introduce an additional 'within-utterance' silence model ('full stop'), because the HDecode decoder assumes an end of turn once 'standard' silence is detected. After building the tied-state triphones, Gaussian mixtures (GM) were split until each state had 8 GMs, with an additional split for silence (16 GMs).

As has been shown in [5], using a domain-specific language model (LM) can vastly increase recognition rate. For our experiments, we built a trigram LM from the archive of the German newspaper *taz (die tageszeitung)*, consisting of 633 611 articles from the years 1986-2000 with 185.9 million words in total. For the 148.5 K LM vocabulary, pronunciations were taken from a semi-automatically generated dictionary. Notably, for lower vocabulary sizes the OOV rate increased significantly (see Table 3). To match the structure of the utterance annotation appropriately with the LM, we replaced the full stops in the articles by a word corresponding to the aforementioned 'within-utterance' silence model. We ended up with an LM size of approx. 400 megabytes in ARPA format, hence the LM fits well into the memory of a state-of-the-art PC, and for our purpose no special effort to cope with memory requirements was needed, unlike in [12]. Table 4 shows the perplexity – measured by the method described in [11] – of the transcription of the training, development, and test sets, and the BCN corpus as a whole, using the unigram, bigram, and trigram LM.

n **words**	10 K	20 K	50 K	100 K	148 K
% OOV	14.1	9.3	4.7	2.1	0.36

Table 3: *OOV rates over the BCN test set when using only the n most frequent words in the texts for building the language model.*

Set	**Unigram**	**Bigram**	**Trigram**
Training	2 487	462	299
Development	2 150	445	291
Test	2 751	473	302
Total	2 474	462	299

Table 4: *Perplexity of the training, development, and test set transcriptions, measured using a language model built from 633 611 articles (185.9 M words) from the German newspaper* taz.

5. Experimental Results

We defined suitable ranges for the decoding parameters on the development set: the word insertion penalty used by HDecode was set to a log probability of -4.0, the HDecode beam width was chosen between 170.0 and 250.0, and values of the LM scale factor between 12.0 and 20.0. Then, we joined the training and development sets and evaluated different combinations of these parameters on the test set. First, Figure 2 shows the WER for the three different feature sets (Section 3) at various beam widths of the decoder (170.0, 180.0, 190.0, 200.0, 225.0, and 250.0). With the minimal pruning (beam width 250.0) and PLP features, a WER of 8.1 % can be achieved. Comparing the results with different feature sets, it can be seen that while a WER of below 10 % is feasible for all three feature sets, the overall picture clearly shows that PLP outperform MFCC and the combination of MFCC and LSP reduced by PCA. The latter exhibit the lowest performance in the test scenario.

Next, Figure 3 shows the trade-off between WER and RTF, again for the three different feature sets. It is remarkable that the 39 MFCC and PLP features are approximately on par concerning RTF, while the 38 PCA features from MFCC and LSP significantly increase the RTF needed for a given WER. Note that in this figure, different RTFs resulted from using the various aforementioned pruning settings, and were computed as the quotient of the total CPU time used by HDecode on the test set and the total utterance length of the test set. The measurement of CPU time was carried out with HTK 3.4 on a Linux PC with a 3.0 GHz AMD Phenom II X4 processor and 8 GB of RAM. Only one of the four cores was used for each decoding process. For the results in Figure 2 and 3, the LM scale factor was 15.0.

Variation of the LM scale factor is another method to fine-tune the trade-off between WER and RTF: Figure 4 depicts the WER induced by different scale factors, while Figure 5 shows the trade-off between WER and RTF in analogy to Figure 3, but varying the LM scale factor instead of the beam width. Here, PLP features were used. Opposing Figure 5 and Figure 3, one can see that variation of the beam width at a constant LM scale factor of 15.0 is more beneficial for improving the trade-off between WER and RTF than keeping the beam width constant (200.0) and varying the LM scale factor. In the latter case, a higher WER of 9.7 % has to be accepted for real-time decoding (with an LM scale factor of 17.0), compared to a WER of about 9.2 % achieved by a beam width of 190.0. Furthermore, a RTF of 0.67 at a WER of 10.82 % can be achieved by a beam width of 170.0, while accomplishing a RTF of 0.79 by a LM scale factor of 20.0 induces a WER of 14.10 %.

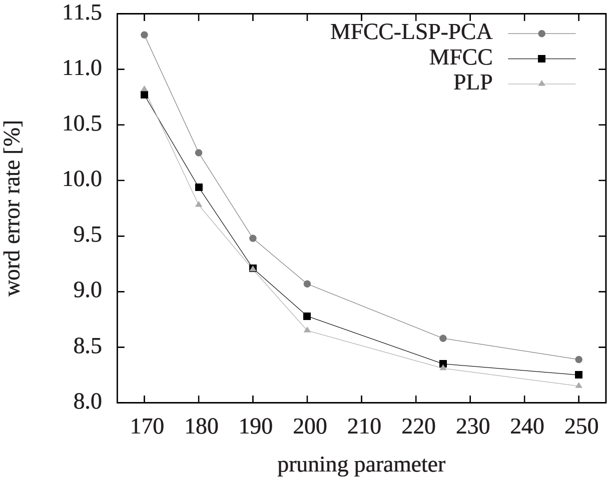

Figure 2: *Effect of beam width on word error rate for the test set of the BCN corpus and three different feature sets.*

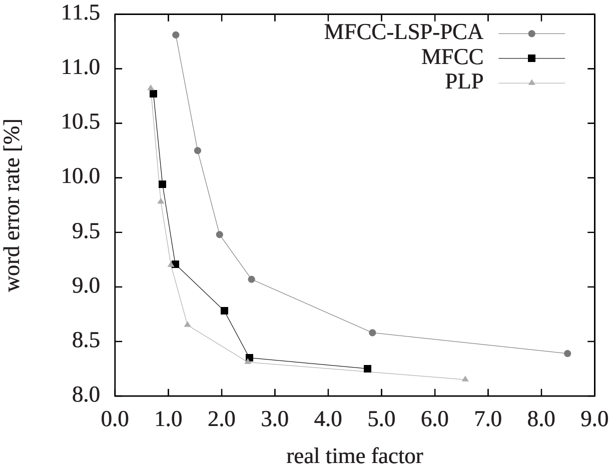

Figure 3: *Trade-off between real-time factor and word error rate for the test set of the BCN corpus when varying the beam width and feature set. The language model scale factor was 15.0. Measurements were carried out on a 3.0 GHz desktop PC.*

As a final note concerning WER, it is noteworthy that sometimes calculation of WER for German speech takes into account equivalence classes for compound words [6]. That is, if the reference transcription contains the word *Sportveranstaltung* (*sports event*) and the decoder splits it up into the sequence *Sport Veranstaltung*, this is *not* considered erroneous (more precisely, as one insertion and one substitution). However, we calculate WER without using such equivalence classes, for two reasons: first, we believe that the decoder should produce an orthographically correct transcription, which *Sport Veranstaltung* is not; second, it is expected that compound words are pronounced without pauses, which the system should be able to detect. The ‘full stop’ word (Section 4) was however ignored in computation of accuracy.

6. Conclusions

We presented a corpus of German broadcast news for fine-tuning and evaluation of German LVCSR systems, focusing on the

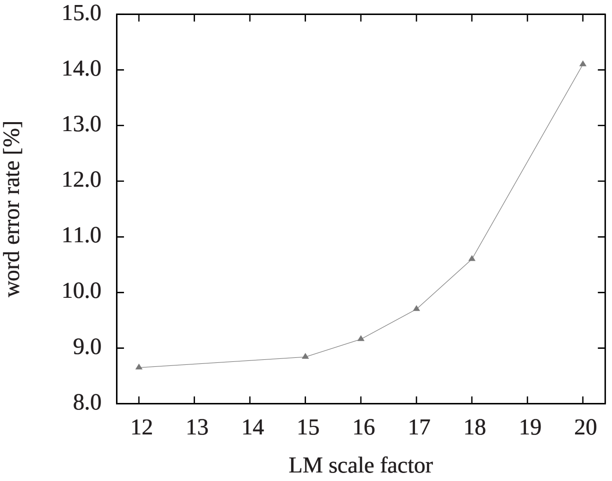

Figure 4: *Effect of language model scale factor on word error rate for the test set of the BCN corpus. A beam width of 200.0 and PLP features were used.*

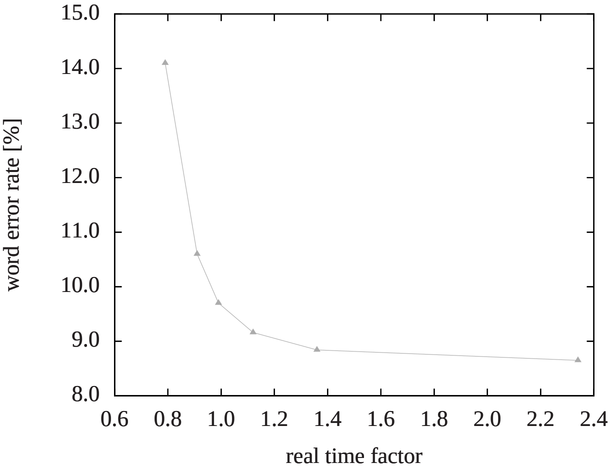

Figure 5: *Trade-off between real-time factor (on a 3.0 GHz PC) and word error rate for the test set of the BCN corpus when varying the language model scale factor at a beam width of 200.0. PLP features were used. Comparing to Figure 3, it can be seen that variation of the beam width generally delivers a lower word error rate for a given real-time factor.*

decoding of clean speech. Using a decoding system based on state-of-the-art LVCSR research and a LM vocabulary size of 148 K, we achieved a word error rate of about 9 % in real-time on a desktop PC. Different feature sets were evaluated, whereby PLP coefficients showed the best performance. Furthermore, we extensively fine-tuned the decoding parameters, and identified the beam widths and LM scale factors needed to achieve various real-time factors. Comparing our results to previous work on broadcast news transcription, we conclude that our results are competitive, especially given that German is a considerably challenging language for LVCSR applications. Naturally, comparisons with previous studies must be done very carefully, as our results were achieved on manually segmented data, and capabilities of PC hardware have developed at a fast pace in the last decade.

In the future, we will use the BCN corpus to evaluate novel hybrid decoders and approaches for feature extraction, e. g. decoders using bidirectional long short-term memory recurrent neural networks or features generated by non-negative matrix factorisation. While both of these have been successfully used in whole-word recognition tasks [13], it will be an interesting research topic to assess their usefulness for German LVCSR.

7. Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

The authors would like to thank Uri Iurgel, Andreas Kosmala, Steffen Werner, and Daniel Willett for their contributions within the ALERT project during their period at the Department of Computer Science at Gerhard-Mercator-University Duisburg.

8. References

[1] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney, M. Pitz, and A. Sixtus, "Large vocabulary continuous speech recognition of broadcast news – the Philips/RWTH approach," *Speech Communication*, vol. 37, pp. 109–131, May 2002.

[2] M. J. F. Gales, D. Kim, P. C. Woodland, H. Chan, D. Mrva, R. Sinha, and S. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1513–1525, 2006.

[3] L. Lamel, A. Messaoudi, and J.-L. Gauvain, "Improved acoustic modeling for transcribing Arabic broadcast data," in *Proc. of Interspeech*, Antwerp, Belgium, August 2007.

[4] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk, "Where are we in transcribing French broadcast news?" in *Proc. of Interspeech*, Lisbon, Portugal, September 2005.

[5] G. Rigoll, "The ALERT-System: Advanced broadcast speech recognition technology for selective dissemination of multimedia information," in *Proc. of ASRU*, Madonna di Campiglio Trento, Italy, December 2001.

[6] K. McTait and M. Adda-Decker, "The 300k LIMSI German broadcast news transcription system," in *Proc. of Eurospeech*, Geneva, Switzerland, September 2003.

[7] H. Yu, Y.-C. Tam, T. Schaaf, S. Stüker, Q. Jin, M. Noamany, and T. Schultz, "The ISL RT04 Mandarin broadcast news evaluation system," in *Proc. of EARS Rich Transcription Workshop (EARS RT)*, Palisades, NY, USA, November 2004.

[8] R. Sinha, M. J. F. Gales, D. Kim, X. Liu, K. Sim, and P. C. Woodland, "The CU-HTK broadcast news transcription system," in *Proc. of ICASSP*, Toulouse, France, May 2006.

[9] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 98–108, 2002.

[10] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - introducing the Munich open-source Emotion and Affect Recognition toolkit," in *Proc. of ACII*, Amsterdam, Netherlands, 2009.

[11] S. J. Young, G. Evermann, M. J. F. Gales, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK book version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[12] D. Willett, C. Neukirchen, and G. Rigoll, "Ducoder - the Duisburg University LVCSR stackdecoder," in *Proc. of ICASSP*, Istanbul, Turkey, June 2000, pp. 1555–1558.

[13] B. Schuller, F. Weninger, M. Wöllmer, Y. Sun, and G. Rigoll, "Non-negative matrix factorization as noise-robust feature extractor for speech recognition," in *Proc. of ICASSP*, Dallas, TX, USA, March 2010, pp. 4562–4565.